DOCUMENT RESUME

ED 425 965 SE 062 136

AUTHOR	Reese, Clyde M.; Jerry, Laura; Ballator, Nada
TITLE	NAEP 1996 Mathematics State Report for New Hampshire. Findings from the National Assessment of Educational Progress.
INSTITUTION	Educational Testing Service, Princeton, NJ.; National Assessment of Educational Progress, Princeton, NJ.
SPONS AGENCY	National Center for Education Statistics (ED), Washington, DC.
REPORT NO	NCES-97-974-NH
PUB DATE	1997-06-00
NOTE	60p.; For overall report, see ED 406 223. For other state reports, see SE 062 110-157. "In collaboration with Peggy Carr, Jeff Haberstroh, Paul Koehler, Phillip Leung, Mary Lindquist, and John Mazzeo."
AVAILABLE FROM	National Library of Education, Office of Educational Research and Improvement, U.S. Department of Education, 555 New Jersey Avenue, NW, Washington, DC 20208-5641; Tel: 800-424-1616 (Toll Free); Web site: http://www.ed.gov/NCES/naep
PUB TYPE	Numerical/Quantitative Data (110) -- Reports - Research (143)
EDRS PRICE	MF01/PC03 Plus Postage.
DESCRIPTORS	Algebra; Functions (Mathematics); Geometry; *Grade 8; Junior High Schools; *Mathematics Achievement; Measurement; *National Competency Tests; Number Concepts; Private Schools; Probability; Problem Solving; Racial Differences; Spatial Ability; Standards; *State Surveys; Statistics; *Student Evaluation; Tables (Data); Test Results
IDENTIFIERS	National Assessment of Educational Progress; *New Hampshire; State Mathematics Assessments

ABSTRACT

The National Assessment of Educational Progress (NAEP) is the only nationally representative and continuing assessment of what students in the United States know and can do in various academic subjects. The 1996 NAEP in mathematics assessed the current level of mathematical achievement as a mechanism for informing education reform. In 1996, 44 states, the District of Columbia, Guam, and the Department of Defense schools took part in the NAEP state mathematics assessment program. The NAEP 1996 state mathematics assessment was at grade 4 and grade 8, although grades 4, 8, and 12 were assessed at the national level. The 1996 state mathematics assessment covered the five content strands: (1) Number Sense, Properties, and Operations; (2) Measurement; (3) Geometry and Spatial Sense; (4) Data Analysis, Statistics, and Probability; and (5) Algebra and Functions. In New Hampshire, 212 students in 12 nonpublic schools were assessed at the eighth-grade level. (To ensure comparability across jurisdictions, the National Center for Education Statistics, has established guidelines for school and student participation rates. In 1996, New Hampshire participated at grade 8 only, with both public and nonpublic school samples. The grade 8 public school sample did not meet the guidelines for publication, due to low participation rate; however, the nonpublic school sample was sufficient to meet the guidelines for publication.) This report describes the mathematics proficiency of New

+++++ ED425965 Has Multi-page SFR---Level=1 +++++

Hampshire eighth-grade nonpublic school students, compares their overall performance to students in the Northeast region of the United States and the entire United States (using data from the NAEP national assessment), presents the average proficiency for the five content strands, and summarizes the performance of subpopulations (gender, race/ethnicity, parents' educational level, Title I participation, and free/reduced lunch program eligibility). To provide a context for the assessment data, participating students, their mathematics teachers, and principals completed questionnaires which focused on: school characteristics (attendance); instructional content (curriculum coverage, standards; amount of homework); delivery of mathematics instruction and its characteristics; use of technology in mathematics instruction; students' own views about mathematics; and conditions facilitating mathematics learning (hours of television watched, parental support, home influences). (ASK)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 425 965

NAEP 1996 MATHEMATICS

State Report for New Hampshire

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

THE NATION'S REPORT CARD

U.S. DEPARTMENT OF EDUCATION OFFICE OF EDUCATIONAL RESEARCH AND IMPROVEMENT

Findings from the National Assessment of Educational Progress

2

What is The Nation's Report Card?

THE NATION'S REPORT CARD, the National Assessment of Educational Progress (NAEP), is the only nationally representative and continuing assessment of what America's students know and can do in various subject areas. Since 1969, assessments have been conducted periodically in reading, mathematics, science, writing, history/geography, and other fields. By making objective information on student performance available to policymakers at the national, state, and local levels, NAEP is an integral part of our nation's evaluation of the condition and progress of education. Only information related to academic achievement is collected under this program. NAEP guarantees the privacy of individual students and their families.

NAEP is a congressionally mandated project of the National Center for Education Statistics, the U.S. Department of Education. The Commissioner of Education Statistics is responsible, by law, for carrying out the NAEP project through competitive awards to qualified organizations. NAEP reports directly to the Commissioner, who is also responsible for providing continuing reviews, including validation studies and solicitation of public comment, on NAEP's conduct and usefulness.

In 1988, Congress established the National Assessment Governing Board (NAGB) to formulate policy guidelines for NAEP. The Board is responsible for selecting the subject areas to be assessed from among those included in the National Education Goals; for setting appropriate student performance levels; for developing assessment objectives and test specifications through a national consensus approach; for designing the assessment methodology; for developing guidelines for reporting and disseminating NAEP results; for developing standards and procedures for interstate, regional, and national comparisons; for determining the appropriateness of test items and ensuring they are free from bias; and for taking actions to improve the form and use of the National Assessment.

The National Assessment Governing Board

Honorable William T. Randall, Chair
Former Commissioner of Education
State of Colorado
Denver, Colorado

Mary R. Blanton, Vice Chair
Attorney
Salisbury, North Carolina

Patsy Cavazos
Principal
W.G. Love Accelerated Elementary School
Houston, Texas

Catherine A. Davidson
Secondary Education Director
Central Kitsap School District
Silverdale, Washington

Edward Donley
Former Chairman
Air Products & Chemicals, Inc.
Allentown, Pennsylvania

Honorable James Edgar
Member Designate
Governor of Illinois
Springfield, Illinois

James E. Ellingson
Fourth-Grade Classroom Teacher
Probstfield Elementary School
Moorhead, Minnesota

Thomas H. Fisher
Director, Student Assessment Services
Florida Department of Education
Tallahassee, Florida

Michael J. Guerra
Executive Director
Secondary Schools Department
National Catholic Educational Association
Washington, DC

Edward H. Haertel
Professor of Education
Stanford University
Stanford, California

Jan B. Loveless
District Communications Specialist
Midland Public Schools
Midland, Michigan

Marilyn McConachie
Former School Board Member
Glenbrook High Schools
Glenview, Illinois

William J. Moloney
Superintendent of Schools
Calvert County Public Schools
Prince Frederick, Maryland

Honorable Annette Morgan
Former Member
Missouri House of Representatives
Jefferson City, Missouri

Mark D. Musick
President
Southern Regional Education Board
Atlanta, Georgia

Mitsugi Nakashima
First Vice-Chairperson
Hawaii State Board of Education
Honolulu, Hawaii

Michael T. Nettles
Professor of Education & Public Policy
University of Michigan
Ann Arbor, Michigan
and Director
Frederick D. Patterson Research Institute
United Negro College Fund

Honorable Norma Paulus
Superintendent of Public Instruction
Oregon State Department of Education
Salem, Oregon

Honorable Roy Romer
Governor of Colorado
Denver, Colorado

Honorable Edgar D. Ross
Judge
Territorial Court of the Virgin Islands
Christiansted, St. Croix
U.S. Virgin Islands

Fannie L. Simmons
Mathematics Coordinator
District 5 of Lexington/Richland County
Ballentine, South Carolina

Adam Urbanski
President
Rochester Teachers Association
Rochester, New York

Deborah Voltz
Assistant Professor
Department of Special Education
University of Louisville
Louisville, Kentucky

Marilyn A. Whirry
Twelfth-Grade English Teacher
Mira Costa High School
Manhattan Beach, California

Dennie Palmer Wolf
Senior Research Associate
Harvard Graduate School of Education
Cambridge, Massachusetts

Ramon C. Cortines (Ex-Officio)
Acting Assistant Secretary
Office of Educational Research
and Improvement
U.S. Department of Education
Washington, DC

Roy Truby
Executive Director, NAGB
Washington, DC

NATIONAL CENTER FOR EDUCATION STATISTICS

NAEP 1996 MATHEMATICS STATE REPORT

for

NEW HAMPSHIRE

Clyde M. Reese
Laura Jerry
Nada Ballator

In collaboration with
Peggy Carr, Jeff Haberstroh
Paul Koehler, Phillip Leung
Mary Lindquist, and John Mazzeo

June 1997

U.S. Department of Education
Office of Educational Research and Improvement

Prepared by Educational Testing Service under a cooperative agreement with the National Center for Education Statistics.

U.S. Department of Education
Richard W. Riley
Secretary

Office of Educational Research and Improvement
Ramon C. Cortines
Acting Assistant Secretary

National Center for Education Statistics
Pascal D. Forgione, Jr.
Commissioner

Education Assessment Group
Gary W. Phillips
Associate Commissioner

June 1997

SUGGESTED CITATION
Reese, C.M., Jerry, L., and Ballator, N.
NAEP 1996 Mathematics State Report for New Hampshire,
Washington, DC: National Center for Education Statistics, 1997.

FOR MORE INFORMATION
Contact:
Arnold A. Goldstein
202-219-1741

For ordering information on this report, write:

National Library of Education
Office of Educational Research and Improvement
U.S. Department of Education
555 New Jersey Avenue, NW
Washington, D.C. 20208-5641

or call 1-800-424-1616 (in the Washington, DC, metropolitan area call 202-219-1651).

This report also is available on the World Wide Web: http://www.ed.gov/NCES/naep

The work upon which this publication is based was performed for the National Center for Education Statistics, Office of Educational Research and Improvement, by Educational Testing Service.

Educational Testing Service is an equal opportunity, affirmative action employer.

Educational Testing Service, *ETS*, and *the ETS logo* are registered trademarks of Educational Testing Service.

Table of Contents

INTRODUCTION 1

OVERVIEW 3

What Is NAEP? 3

What Was Assessed? 4

Who Was Assessed? 5

TABLE 1 Profile of Students in New Hampshire, Northeast Region, and Nation 6

TABLE 2 Profile of the Population Assessed in New Hampshire 9

RESULTS FOR NONPUBLIC SCHOOLS 11

Interpreting NAEP Mathematics Results 11

TABLE 3 Distribution of Mathematics Scale Scores for Students in Nonpublic Schools 12

FIGURE 1 Policy Definitions of NAEP Achievement Levels 13

TABLE 4 Percentage of Students in Nonpublic Schools Attaining Mathematics Achievement Levels 14

FIGURE 2 Mathematics Achievement Levels 15

APPENDIX A Reporting the NAEP 1996 Mathematics Results 19

APPENDIX B NAEP 1996 Mathematics Assessment 35

APPENDIX C Technical Appendix 39

APPENDIX D Setting the Achievement Levels 53

INTRODUCTION

New Hampshire was one of the original participants in the state-level National Assessment of Educational Progress (NAEP) in 1990. Results were reported for New Hampshire in 1990 (for mathematics at grade 8), in 1992 (for mathematics at grades 4 and 8, and reading at grade 4), and again in 1994 (for reading at grade 4). In 1994, New Hampshire participated with both public and nonpublic school samples, but met the participation rate requirements for publication of results only for their public school sample.

To ensure comparability across jurisdictions, NCES has established guidelines for school and student participation rates. Appendix A highlights these guidelines, which are applied separately for public and nonpublic schools. For jurisdictions failing to meet the initial school participation rate of 70 percent for either public or nonpublic schools, appropriate results are not reported. Jurisdictions that exceed the 70 percent rate but fail to meet others of these guidelines are noted in tables and figures in NAEP reports containing state-by-state results.

In 1996, New Hampshire again participated at grade 8 only, but with both public and nonpublic school samples. The grade 8 public school sample did not meet the guidelines for publication, due to low participation rate (see Appendix A); however, the nonpublic school sample was sufficient to meet the guidelines for publication.

The results of the NAEP 1996 mathematics assessment in New Hampshire's nonpublic schools are presented here. The total sample size for nonpublic schools is modest, and there are a small number of results to report. Only those results based on pre-established NAEP minimum sample sizes are reported. The results that can be reported are students' average scale scores and percentages of students reaching the *Basic*, *Proficient*, and *Advanced* achievement levels. Also included are tables showing the demographic composition of the sample, and the participation rates for sample components.

A full set of Appendices is included:

Appendix A	Reporting the NAEP 1996 Mathematics Results
Appendix B	NAEP 1996 Mathematics Assessment
Appendix C	Technical Appendix
Appendix D	Setting the Achievement Levels

OVERVIEW

Monitoring the performance of students in subjects such as mathematics is a key concern of the citizens, policy makers, and educators involved with educational reform efforts. The 1996 National Assessment of Educational Progress (NAEP) in mathematics (as well as the two previous NAEP assessments in mathematics in 1990 and 1992) assessed the current level of mathematics achievement as a mechanism for informing education reform. In 1996, New Hampshire participated in NAEP at grade 8 but only the nonpublic schools met the participation guidelines. This report contains those results.

What Is NAEP?

The National Assessment of Educational Progress (NAEP) is the only nationally representative and continuing assessment of what students in the United States know and can do in various academic subjects. NAEP is authorized by Congress and directed by the National Center of Education Statistics of the U.S. Department of Education. The National Assessment Governing Board (NAGB), an independent body, provides policy guidance for NAEP.

Since its inception in 1969, NAEP's mission has been to collect, analyze, and produce valid and reliable information about the academic performance of students in the United States in various learning areas. In 1990, the mission of NAEP was expanded to provide state-by-state results on academic achievement. Participation in the state-by-state NAEP is voluntary and has grown from 40 states and territories in 1990 to 48 in 1996.

NAEP has also become a valuable tool in tracking progress towards the National Education Goals. The subjects assessed by NAEP are those highlighted at the 1989 Education Summit and later legislation.[1] The NAEP 1996 assessment in mathematics marks the third time the subject has been assessed with the new framework in the 1990s, enabling policy makers and educators to track mathematics achievement since the release of the National Council of Teachers of Mathematics (NCTM) *Curriculum and Evaluation Standards for School Mathematics*[2] in 1989.

[1] Executive Office of the President. *National Goals for Education.* (Washington, DC: Government Printing Office, 1990); Goals 2000: Educate America Act, Pub. L. No. 103-227 (1994).

[2] National Council of Teachers of Mathematics. *Curriculum and Evaluation Standards for School Mathematics.* (Reston, VA: NCTM, 1989).

What Was Assessed?

The NAEP assessment measures a mathematics domain containing five mathematics strands (number sense, properties, and operations; measurement; geometry and spatial sense; data analysis, statistics, and probability; and algebra and functions). Questions involving content from one or more of the strands are also categorized according to the domains of mathematical abilities and mathematical power. The first of these, mathematical abilities, describes the nature of the knowledge or processes involved in successfully handling the task presented by the question. It may reflect conceptual understanding, procedural knowledge, or a combination of both in problem solving. The second domain, mathematical power, reflects processes stressed as major goals of the mathematical curriculum. Mathematical power refers to the students' ability to reason, to communicate, and to make connections of concepts and skills across mathematical strands, or from mathematics to other curricular areas.

The mathematics framework for the NAEP 1996 assessment is a revision of that used in the 1990 and 1992 assessments. Changes were made to the earlier framework in light of the NCTM Standards and changes taking place in school mathematics programs. The previous NAEP mathematics framework was refined and sharpened so that the 1996 assessment would: (1) more adequately reflect recent curricular emphases and objects and yet (2) maintain a connection with the 1990 and 1992 assessments to measure trends in student performance. Prior to the 1996 assessment, investigations were conducted to ensure that results from the assessment could be reported on the existing NAEP mathematics scale. The conclusion drawn from these investigations was that results from the 1990, 1992, and 1996 assessments could be reported on a common scale and trends in mathematics performance since 1990 examined. Appendix B briefly highlights selected changes in the current NAEP mathematics framework.

The conception of mathematical power as reasoning, connections, and communication has played an increasingly important role in measuring student achievement. In 1990, the NAEP assessment included short constructed-response questions as a way to begin addressing mathematical communication. In 1992, the extended constructed-response questions included on the assessment required students not only to communicate their ideas but also to demonstrate the reasoning they used to solve problems. The 1996 assessment continued to emphasize mathematical power by including constructed-response questions focusing on reasoning and communication and by requiring students to connect their learning across mathematical content strands. These connections were addressed within individual questions reaching across content strands and by families of questions contained within a single content strand.

In real life, few mathematical situations can be clearly classified as belonging to one content strand or another, and few situations require only one fact of mathematics thinking. Therefore, many of the questions are classified in a number of ways. In addition to being classified by all applicable content strands, each question was classified by its assessment of applicable mathematical abilities (procedural knowledge, conceptual understanding, and problem solving) and mathematical powers (reasoning, communication, and connections). The content strands, mathematical abilities, and mathematical power combine to form the framework for the NAEP assessment. (A brief description of the five content strands is presented in Appendix B.)

The framework continued the shift from multiple-choice questions to questions that required students to construct responses. in 1996, more than 50 percent of student assessment time was devoted to constructed-response questions. Two types of constructed-response questions were included — (1) short constructed-response questions that required students to provide answers to computation problems or to describe solutions in one or two sentences, and (2) extended constructed-response questions that required students to provide longer responses when answering the questions.

Who Was Assessed?

Eighth-Grade School and Student Characteristics

Table 1 provides a profile of the demographic characteristics of the eighth-grade students in nonpublic schools in New Hampshire, the Northeast region, and the nation. This report contains assessment results for nonpublic school students only. New Hampshire participated in the NAEP mathematics assessment in 1990 and 1992 and met the minimum guidelines for publication for their public school results in both years. For the 1996 NAEP, although both public and nonpublic schools participated, only nonpublic schools met the guidelines for publication of results. As described in Appendix A, the state data and the regional and national data are drawn from separate samples.

THE NATION'S REPORT CARD naep 1996 State Assessment

TABLE 1 — GRADE 8

Profile of Students in New Hampshire, the Northeast Region, and the Nation

Demographic Subgroups		Nonpublic Schools
		Percentage
RACE/ETHNICITY		
New Hampshire	White	93 (2.6)
	Black	2 (0.8)
	Hispanic	4 (1.7)
	Asian/Pacific Islander	1 (****)
	American Indian	0 (****)
Northeast	White	79 (8.4)
	Black	11 (****)
	Hispanic	8 (1.8)
	Asian/Pacific Islander	1 (0.5)
	American Indian	1 (****)
Nation	White	80 (3.4)
	Black	7 (2.5)
	Hispanic	9 (2.0)
	Asian/Pacific Islander	3 (0.7)
	American Indian	1 (0.3)
TYPE OF LOCATION*		
New Hampshire	Central city	45 (16.2)
	Urban fringe/Large town	9 (****)
	Rural/Small town	46 (13.8)
Nation	Central city	69 (6.8)
	Urban fringe/Large town	22 (6.1)
	Rural/Small town	9 (4.2)
PARENTS' EDUCATION		
New Hampshire	Did not finish high school	1 (****)
	Graduated from high school	10 (4.9)
	Some education after high school	12 (3.3)
	Graduated from college	74 (6.5)
	I don't know	4 (1.2)
Northeast	Did not finish high school	2 (0.7)
	Graduated from high school	15 (3.8)
	Some education after high school	16 (1.8)
	Graduated from college	56 (8.1)
	I don't know	11 (4.4)
Nation	Did not finish high school	2 (0.4)
	Graduated from high school	13 (1.5)
	Some education after high school	16 (0.9)
	Graduate from college	61 (2.7)
	I don't know	9 (1.6)
GENDER		
New Hampshire	Male	53 (9.5)
	Female	47 (9.5)
Northeast	Male	56 (6.0)
	Female	44 (6.0)
Nation	Male	53 (2.1)
	Female	47 (2.1)

THE NATION'S REPORT CARD 1996 State Assessment

TABLE 1 — GRADE 8 (continued)

Profile of Students in New Hampshire, the Northeast Region, and the Nation

Demographic Subgroups		Nonpublic Schools
		Percentage
TITLE 1		
New Hampshire	Participated	2 (****)
	Did not participate	98 (****)
Northeast	Participated	2 (****)
	Did not participate	98 (****)
Nation	Participated	2 (1.2)
	Did not participate	98 (1.2)
FREE/REDUCED-PRICE LUNCH		
New Hampshire	Eligible	1 (****)
	Not eligible	49 (17.5)
	Information not available	50 (17.4)
Northeast	Eligible	4 (****)
	Not eligible	47 (16.7)
	Information not available	48 (16.8)
Nation	Eligible	4 (1.5)
	Not eligible	53 (7.5)
	Information not available	43 (7.6)

The standard errors of the statistics appear in parentheses. It can be said with about 95 percent confidence that, for each population of interest, the value for the entire population is within ± 2 standard errors of the estimate for the sample. In comparing two estimates, one must use the standard error of the difference (see Appendix A for details). The percentages for Race/Ethnicity may not add to 100 percent because some students categorized themselves as "Other." * Characteristics of the school sample do not permit reliable regional results for type of location. **** Standard error estimates cannot be accurately determined.
SOURCE: National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1996 Mathematics Assessment.

Schools and Students Assessed

For public schools in 1996, the weighted school participation rate before substitution was 69 percent for New Hampshire; a level of 70 percent was required for publication of public school results. Table 2 summarizes participation data for nonpublic schools and students sampled in New Hampshire for the 1996 state assessment program in mathematics.[3] Also included is the participation profile for public schools in 1996.

In New Hampshire, 12 nonpublic schools participated in the 1996 eighth-grade mathematics assessment. The weighted school participation rate after substitution in 1996 was 85 percent for nonpublic schools, which means that the eighth-grade students in this sample were directly representative of 85 percent of all the eighth-grade public school students in New Hampshire.

In New Hampshire 212 nonpublic school eighth-grade students were assessed in 1996. The weighted student participation rate was 96 percent for nonpublic schools. This means that the sample of eighth-grade students who took part in the assessment was directly representative of 96 percent of the eligible nonpublic school student population in participating schools in New Hampshire (that is, all students from the population represented by the participating schools, minus those students excluded from the assessment). The overall weighted response rate (school rate times student rate) was 82 percent for nonpublic schools. This means that the sample of students who participated in the assessment was directly representative of 82 percent of the eligible eighth-grade nonpublic school population in New Hampshire.

In each school, a random sample of students was selected to participate in the assessment. In 1996, on the basis of sample estimates, 1 percent of the eighth-grade nonpublic school population in New Hampshire was classified as having limited English proficiency (LEP). In nonpublic schools at the eighth grade, 3 percent of the students had an Individualized Education Plan (IEP). An IEP is a plan written for a student who has been determined to be eligible for special education. The IEP typically sets forth goals and objectives for the student and describes a program of activities and/or related services necessary to achieve the goals and objectives.

[3] For a detailed discussion of the NCES guidelines for sample participation, see Appendix A of this report or the *Technical Report of the NAEP 1996 State Assessment Program in Mathematics*. (Washington, DC: National Center for Education Statistics, 1997).

THE NATION'S REPORT CARD 1996 State Assessment

TABLE 2 — GRADE 8

Profile of the Population Assessed in New Hampshire

	1996
	Nonpublic Schools
SCHOOL PARTICIPATION	
Weighted school participation rate before substitution	85%
Weighted school participation rate after substitution	85%
Number of schools originally sampled	19
Number of schools not eligible	4
Number of schools in original sample participating	12
Number of substitute schools provided	2
Number of substitute schools participating	0
Total number of participating schools	12
STUDENT PARTICIPATION	
Weighted student participation rate after makeups	96%
Number of students selected to participate in the assessment	224
Number of students withdrawn from the assessment	4
Percentage of students who were of Limited English Proficiency	1%
Percentage of students excluded from the assessment due to Limited English Proficiency	1%
Percentage of students who had an Individualized Education Plan	3%
Percentage of students excluded from the assessment due to Individualized Education Plan status	0%
Number of students to be assessed	219
Number of students assessed	212
Overall weighted response rate	82%

RESULTS FOR NONPUBLIC SCHOOLS

NAEP Mathematics Results for New Hampshire

The NAEP 1996 state assessment program in mathematics provides a wealth of information on the mathematical abilities and skills of the fourth-and eighth-grade students in participating jurisdictions. To maximize usefulness to policy makers, educators, parents, and other interested parties, the NAEP results are presented both as average scale scores on the NAEP mathematics scale (Table 3) and in terms of the percentage of students attaining NAEP mathematics achievement levels (Table 4). Thus, NAEP results not only provide information about what students *know and can do*, but also indicate whether their achievement meets expectations of what students *should know and should be able to do*. Furthermore, the descriptions of skills and abilities expected of students at each achievement level help make the reporting of assessment results more meaningful.

Interpreting NAEP Results

This report describes mathematics performance for eighth graders in New Hampshire nonpublic schools and compares the results with those of eighth grade students in nonpublic schools in the Northeast region and in the nation.

Because the percentages of students and their average mathematics scale scores are based on samples — rather than on the entire population of eighth graders in a jurisdiction — the numbers reported are necessarily *estimates*. As such, they are subject to a measure of uncertainty, reflected in the *standard error* of the estimate. When the percentages or average scale scores of certain groups are compared, it is essential to take the standard error into account, rather than to rely solely on observed similarities or differences. Therefore, the comparisons discussed in this report are based on *statistical tests* that consider both the magnitude of the difference between the means or percentages and the standard errors of those statistics.

The statistical tests determine whether the evidence — based on the data from the groups in the *sample* — is strong enough to conclude that the averages or percentages are really different for those groups in the *population*. If the evidence is strong (i.e., the difference is statistically significant), the report describes the group averages or percentages as being different (e.g., the states' students performed *higher than* or *lower than* the nation's students) — regardless of whether the sample averages or sample percentages appear to be about the same or not. If the evidence is not sufficiently strong (i.e., the difference is not statistically significant), the averages or percentages are described as being *not significantly different* — again, regardless of whether

the sample averages or sample percentages appear to be about the same or widely discrepant. The reader is cautioned to rely on the results of the statistical tests rather than on the apparent magnitude of the difference between sample averages or percentages to determine whether those sample differences are likely to represent actual differences between the groups in the population. The statistical tests are discussed in greater detail in Appendix A.

The Mathematics Scale

Students' responses to the NAEP 1996 mathematics assessment were analyzed to determine the percentage of students responding correctly to each multiple-choice question and the percentage of students responding in each of several score categories for constructed-response questions. Item response theory (IRT) methods were used to produce across-grade scales that summarized results for each of the five mathematics content strands discussed earlier. Each of the content-strand scales, which range from 0 to 500, was linked to its corresponding scale from 1990 and 1992 through IRT equating.

An overall composite scale was developed by weighting the separate content-strand scales based on the relative importance to each content strand in the NAEP mathematics framework. The resulting scale, which was also linked to the 1990 and 1992 mathematics composite scales, is the reporting metric used to present results. (Details of the scaling procedures are presented in the *NAEP 1996 Technical Report* and in the *Technical Report of the NAEP 1996 State Assessment Program in Mathematics.*)

THE NATION'S REPORT CARD naep 1996 State Assessment

TABLE 3 — GRADE 8

Distribution of Mathematics Scale Scores for Students in Nonpublic Schools

	Average Scale Score	10th Percentile	25th Percentile	50th Percentile	75th Percentile	90th Percentile
Nonpublic Schools						
1996 New Hampshire	293 (4.3)	257 (5.2)	274 (4.0)	293 (3.5)	313 (5.9)	329 (4.7)
Northeast	281 (5.1)!	243 (12.4)!	260 (6.3)!	283 (6.4)!	302 (5.9)!	320 (8.0)!
Nation	284 (2.4)	242 (4.4)	262 (4.6)	286 (1.8)	307 (2.4)	326 (3.1)

The NAEP mathematics scale ranges from 0 to 500. The standard errors of the statistics appear in parentheses. It can be said with about 95 percent confidence that, for each population of interest, the value for the entire population is within ± 2 standard errors of the estimate for the sample. In comparing two estimates, one must use the standard error of the difference (see Appendix A for details). ! Interpret with caution — the nature of the sample does not allow accurate determination of the variability of this statistic.
SOURCE: National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1996 Mathematics Assessment.

- In New Hampshire, the average mathematics scale score of students attending nonpublic schools (293) was not significantly different from* that of nonpublic school students across the nation (284).

* Although the difference may appear large, recall that "significance" here refers to "statistical significance."

Mathematics Achievement Levels

Results for the NAEP 1996 assessment in mathematics are also reported using the mathematics achievement levels that were authorized by the NAEP legislation and adopted by the National Assessment Governing Board. The achievement levels are based on collective judgments about what students *should know and be able to do* relative to the body of content reflected in the NAEP mathematics assessment. Three levels were defined for each grade — *Basic, Proficient,* and *Advanced.* The levels were defined by a broadly representative panel of teachers, education specialists, and members of the general public.

For reporting purposes, the achievement levels for each grade are placed on the NAEP mathematics scale. Figure 1 presents the policy definitions of the achievement levels, while Figure 2 contains specific descriptions for the levels at grade 8.

Figure 1. Policy Definitions of NAEP Achievement Levels

Basic	This level denotes partial mastery of prerequisite knowledge and skills that are fundamental for proficient work at each grade.
Proficient	This level represents solid academic performance for each grade assessed. Students reaching this level have demonstrated competency over challenging subject matter, including subject-matter knowledge, application of such knowledge to real-world situations, and analytical skills appropriate to the subject matter.
Advanced	This level signifies superior performance.

It should be noted that setting achievement levels is a relatively new process for NAEP, and it is still in transition. Some evaluations have concluded that the percentage of students at certain levels may be underestimated.[4] On the other hand, critiques of those evaluations have asserted that the weight of the empirical evidence does not support such conclusions.[5] A further review is currently being conducted by the National Academy of Sciences.

[4] General Accounting Office. *Educational Achievement Standards: NAGB's Approach Yields Misleading Interpretations.* (Washington, DC: General Accounting Office, 1993); National Academy of Education. *Setting Performance Standards for Student Achievement.* A report of the National Academy of Education Panel on the evaluation of the NAEP Trial State Assessment: An evaluation of the 1992 achievement levels. (Stanford, CA: National Academy of Education, 1993).

[5] Cizek, G. *Reactions to the National Academy of Education report.* (Washington, DC: National Assessment Governing Board, 1993); Kane, M. *Comments on the NAE evaluation of the NAGB achievement levels.* (Washington, DC: National Assessment Governing Board, 1993); *NAEP Reading Revisited: An Evaluation of the 1992 Achievement Levels Descriptions.* (American College Testing, Washington, DC: National Assessment Governing Board, 1993); *Technical Report on Setting Achievement Levels on the 1992 National Assessment of Educational Progress in Mathematics, Reading, and Writing.* (American College Testing, Washington, DC: National Assessment Governing Board, 1993).

The student achievement levels in this report have been developed carefully and responsibly, and the procedures used have been refined and revised as new technologies have become available. Upon review of the available information, the Commissioner of Education Statistics has judged that the achievement levels are in a developmental status. However, the Commissioner and the Governing Board also believe that the achievement levels are useful and valuable for reporting on the educational achievement of students in the United States.

THE NATION'S REPORT CARD 1996 State Assessment

TABLE 4 — GRADE 8

Percentage of Students in Nonpublic Schools Attaining Mathematics Achievement Levels

		Advanced	At or Above Proficient	At or Above Basic	Below Basic
Nonpublic Schools					
1996	New Hampshire	8 (3.3)	43 (6.5)	86 (3.2)	14 (3.2)
	Northeast	3 (2.1)	28 (5.6)	73 (7.0)	27 (7.0)
	Nation	6 (1.2)	33 (2.9)	75 (2.8)	25 (2.8)

The standard errors of the statistics appear in parentheses. It can be said with about 95 percent confidence that, for each population of interest, the value for the entire population is within ± 2 standard errors of the estimate for the sample. In comparing two estimates, one must use the standard error of the difference (see Appendix A for details).
SOURCE: National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1996 Mathematics Assessment.

- The percentage of nonpublic school students in New Hampshire who performed at or above the *Proficient* level (43 percent) was not significantly different from* that of nonpublic school students across the nation (33 percent).

Description of Mathematics Achievement Levels

The three mathematics achievement levels for grade 8 are described in terms specific to the mathematics assessment in Figure 2. Examples of questions appropriate at each achievement level are also provided. It should be noted that constructed-response questions occur in the assessment at all levels of mathematics achievement.

* Although the difference may appear large, recall that "significance" here refers to "statistical significance."

THE NATION'S REPORT CARD 1996 State Assessment

FIGURE 2

Mathematics Achievement Levels

GRADE 8

NAEP mathematics content strands: (1) Number Sense, Properties, and Operations; (2) Measurement; (3) Geometry and Spatial Sense; (4) Data Analysis, Statistics, and Probability; (5) Algebra and Functions.

Skills are cumulative across all levels — from *Basic* to *Proficient* to *Advanced*.

BASIC LEVEL	**Eighth-grade students performing at the *Basic* level *should exhibit evidence of conceptual and procedural understanding in the five NAEP content strands*. This level of performance signifies an understanding of arithmetic operations — including estimation — on whole numbers, decimals, fractions, and percents. In relation to the NAEP mathematics scale, *Basic*-level achievement for eighth grade is defined by scale scores at or above 262.**

Specifically, eighth graders performing at the *Basic* level should complete problems correctly with the help of structural prompts such as diagrams, charts, and graphs. They should be able to solve problems in all NAEP content strands through the appropriate selection and use of strategies and technological tools — including calculators, computers, and geometric shapes. Students at this level should also be able to use fundamental algebraic and informal geometric concepts in problem solving.

As they approach the *Proficient* level, students at the *Basic* level should be able to determine which of available data are necessary and sufficient for correct solutions and use them in problem solving. However, these eighth graders may show limited skill in communicating mathematically.

PROFICIENT LEVEL	**Eighth-grade students performing at the *Proficient* level *should apply mathematical concepts and procedures consistently to complex problems in the five NAEP content strands*. In relation to the NAEP mathematics scale, *Proficient*-level achievement for eighth grade is defined by scale scores at or above 299.**

Specifically, eighth graders performing at the *Proficient* level should be able to conjecture, defend their ideas, and give supporting examples. They should understand the connections between fractions, percents, decimals, and other mathematical topics such as algebra and functions. Students at the *Proficient* level are expected to have a thorough understanding of basic level arithmetic operations — an understanding sufficient for problem solving in practical situations.

Quantity and spatial relationships in problem solving and reasoning should be familiar to them, and they should be able to convey underlying reasoning skills beyond the level of arithmetic. They should be able to compare and contrast mathematical ideas and generate their own examples. These students should make inferences from data and graphs; apply properties of informal geometry; and accurately use the tools of technology. Students at this level should understand the process of gathering and organizing data and be able to calculate, evaluate, and communicate results within the domain of statistics and probability.

ADVANCED LEVEL	**Eighth-grade students at the *Advanced* level *should be able to reach beyond the recognition, identification, and application of mathematical rules in order to generalize and synthesize concepts and principles in the five NAEP content strands*. In relation to the NAEP mathematics scale, *Advanced*-level achievement for eighth grade is defined by scale scores at or above 333.**

Specifically, eighth graders performing at the *Advanced* level should be able to probe examples and counterexamples in order to shape generalizations from which they can develop models. Eighth graders performing at this level should use number sense and geometric awareness to consider the reasonableness of an answer. They are expected to use abstract thinking to create unique problem-solving techniques and explain the reasoning processes underlying their conclusions.

THE NATION'S REPORT CARD 1996 State Assessment

FIGURE 2 (continued)

Mathematics Achievement Levels

Grade 8 Basic-Level Example Item

Which of the following is both a multiple of 3 and a multiple of 7?

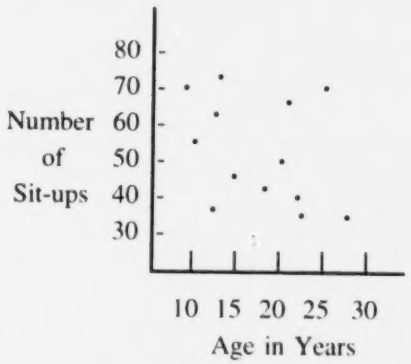
A. 7,007
B. 8,192
*C. 21,567
D. 22,287
E. 40,040

1992 Percent Correct	
Nation	76 (1.3)

Did you use the calculator on this question?

Yes No

Grade 8 Proficient-Level Example Item



In the graph above, each dot shows the number of sit-ups and the corresponding age for one of 13 people. According to this graph, what is the median number of sit-ups for these 13 people?

A. 15
B. 20
C. 45
*D. 50
E. 55

1992 Percent Correct	
Nation	23 (1.4)

Did you use the calculator on this question?

Yes No

20

THE NATION'S REPORT CARD 1996 State Assessment

FIGURE 2 (continued)

Mathematics Achievement Levels

Grade 8 Advanced-Level Example Item

A	*B*
2	5
4	9
6	13
8	17
.	.
.	.
.	.
14	?

If the pattern shown in the table were continued, what number would appear in the box at the bottom of column *B* next to 14?

A. 19
B. 21
C. 23
D. 25
*E. 29

1992 Percent Correct	
Nation	25 (1.4)

APPENDIX A

Reporting NAEP 1996 Mathematics Results for New Hampshire

A.1 Participation Guidelines

As was discussed in the Introduction, unless the overall participation rate is sufficiently high for a jurisdiction, there is a risk that the assessment results for that jurisdiction will be subject to appreciable nonresponse bias. Moreover, even if the overall participation rate is high, there may be significant nonresponse bias if the nonparticipation that does occur is heavily concentrated among certain types of schools or students. The following guidelines concerning school and student participation rates in the state assessment program were established to address four significant ways in which nonresponse bias could be introduced into the jurisdiction sample estimates. The guidelines determining a jurisdiction's eligibility to have its results published are presented below. Also presented below are the conditions that will result in a jurisdiction's receiving a notation in the 1996 reports. Note that in order for a jurisdiction's results to be published with no notations, that jurisdiction must satisfy all guidelines. (A more complete discussion of the NAEP participation guidelines can be found in the *Technical Report of the NAEP 1996 State Assessment Program in Mathematics*.)

Guidelines on the Publication of NAEP Results

Guideline 1 — Publication of Public School Results

A jurisdiction will have its public school results published in the *NAEP 1996 Mathematics Report Card* (or in other reports that include all state-level results) if and only if its weighted participation rate for the initial sample of public schools is greater than or equal to 70 percent. Similarly, a jurisdiction will receive a separate *NAEP 1996 Mathematics State Report* if and only if its weighted participation rate for the initial sample of public schools is greater than or equal to 70 percent.

Guideline 2 — Publication of Nonpublic School Results

A jurisdiction will have its nonpublic school results published in the *NAEP 1996 Mathematics Report Card* (or in other reports that include all state-level results) if and only if its weighted participation rate for the initial sample of nonpublic schools is greater than or equal to 70 percent **AND** meets minimum sample size requirements.[1] A jurisdiction eligible to receive a separate *NAEP 1996 Mathematics State Report* under guideline 1 will have its nonpublic school results included in that report if and only if that jurisdiction's weighted participation rate for the initial sample of nonpublic schools is greater than or equal to 70 percent **AND** meets minimum sample size requirements. If a jurisdiction meets guideline 2 but fails to meet guideline 1, a separate *NAEP 1996 Mathematics State Report* will be produced containing only nonpublic school results.

Guideline 3 — Publication of Combined Public and Nonpublic School Results

A jurisdiction will have its combined results published in the *NAEP 1996 Mathematics Report Card* (or in other reports that include all state-level results) if and only if both guidelines 1 and 2 are satisfied. Similarly, a jurisdiction eligible to receive a separate *NAEP 1996 Mathematics State Report* under guideline 1 will have its combined results included in that report if and only if guideline 2 is also met.

Guidelines for Notations of NAEP Results

Guideline 4 — Notation for Overall Public School Participation Rate

A jurisdiction that meets guideline 1 will receive a notation if its weighted participation rate for the initial sample of public schools was below 85 percent **AND** the weighted public school participation rate after substitution was below 90 percent.

Guideline 5 — Notation for Overall Nonpublic School Participation Rate

A jurisdiction that meets guideline 2 will receive a notation if its weighted participation rate for the initial sample of nonpublic schools was below 85 percent **AND** the weighted nonpublic school participation rate after substitution was below 90 percent.

[1] Minimum participation size requirements for reporting nonpublic school data consist of two components: (1) a school sample size of six or more participating schools and (2) an assessed student sample size of at least 62.

Guideline 6 — Notation for Strata-Specific Public School Participation Rate

A jurisdiction that is not already receiving a notation under guideline 4 will receive a notation if the sample of public schools included a class of schools with similar characteristics that had a weighted participation rate (after substitution) of below 80 percent, and from which the nonparticipating schools together accounted for more than five percent of the jurisdiction's total weighted sample of public schools. The classes of schools from each of which a jurisdiction needed minimum school participation levels were determined by degree of urbanization, minority enrollment, and median household income of the area in which the school is located.

Guideline 7 — Notation for Strata-Specific Nonpublic School Participation Rate

A jurisdiction that is not already receiving a notation under guideline 5 will receive a notation if the sample of nonpublic schools included a class of schools with similar characteristics that had a weighted participation rate (after substitution) of below 80 percent, and from which the nonparticipating schools together accounted for more than five percent of the jurisdiction's total weighted sample of nonpublic schools. The classes of schools from each of which a jurisdiction needed minimum school participation levels were determined by type of nonpublic school (Catholic versus non-Catholic) and location (metropolitan versus nonmetropolitan).

Guideline 8 — Notation for Overall Student Participation Rate in Public Schools

A jurisdiction that meets guideline 1 will receive a notation if the weighted student response rate within participating public schools was below 85 percent.

Guideline 9 — Notation for Overall Student Participation Rate in Nonpublic Schools

A jurisdiction that meets guideline 2 will receive a notation if the weighted student response rate within participating nonpublic schools was below 85 percent.

Guideline 10 — Notation for Strata-Specific Student Participation Rates in Public Schools

A jurisdiction that is not already receiving a notation under guideline 8 will receive a notation if the sampled students within participating public schools included a class of students with similar characteristics that had a weighted student response rate of below 80 percent, and from which the nonresponding students together accounted for more than five percent of the jurisdiction's weighted assessable public school student sample. Student groups from which a jurisdiction needed minimum levels of participation were determined by the age of the student, whether or not the student was classified as a student with a disability (SD) or of limited English proficiency (LEP), and the type of assessment session (monitored or unmonitored), as well as school level of urbanization, minority enrollment, and median household income of the area in which the school is located.

Guideline 11 — Notation for Strata-Specific Student Participation Rates in Nonpublic Schools

A jurisdiction that is not already receiving a notation under guideline 9 will receive a notation if the sampled students within participating nonpublic schools included a class of students with similar characteristics that had a weighted student response rate of below 80 percent, and from which the nonresponding students together accounted for more than five percent of the jurisdiction's weighted assessable nonpublic school student sample. Student groups from which a jurisdiction needed minimum levels of participation were determined by the age of the student, whether or not the student was classified as a student with a disability (SD) or of limited English proficiency (LEP), and the type of assessment session (monitored or unmonitored), as well as type and location of school.

A.2 NAEP Reporting Groups

The state assessment program provides results for groups of students defined by shared characteristics — region of the country, gender, race/ethnicity, parental education, location of the school, type of school, participation in Title I programs, and eligibility for the free/reduced-price lunch component of the National School Lunch Program. Based on criteria described later in this appendix, results are reported for subpopulations only when sufficient numbers of students and adequate school representation are present. For public school students, the minimum requirement is at least 62 students in a particular subgroup from at least 5 primary sampling units (PSUs).[2] For nonpublic school students, the minimum requirement is 62 students from at least 6 different schools for the state assessment program or from at least 5 PSUs for the national assessment. However, the data for all students, regardless of whether their subgroup was reported separately, were included in computing overall results. Definitions of the subpopulations referred to in this report are presented on the following pages.

Region

Results are reported for four regions of the nation: Northeast, Southeast, Central, and West. States included in each region are shown in Figure A.1. All 50 states and the District of Columbia are listed. Territories and the two Department of Defense Educational Activities jurisdictions were not assigned to any region.

Regional results are based on national assessment samples, not on aggregated state assessment program samples. Thus, the regional results are based on a sample that is different and separate from that used to report the state results.

THE NATION'S REPORT CARD 1996 State Assessment

FIGURE A.1

Regions of the Country

NORTHEAST	SOUTHEAST	CENTRAL	WEST
Connecticut	Alabama	Illinois	Alaska
Delaware	Arkansas	Indiana	Arizona
District of Columbia	Florida	Iowa	California
Maine	Georgia	Kansas	Colorado
Maryland	Kentucky	Michigan	Hawaii
Massachusetts	Louisiana	Minnesota	Idaho
New Hampshire	Mississippi	Missouri	Montana
New Jersey	North Carolina	Nebraska	Nevada
New York	South Carolina	North Dakota	New Mexico
Pennsylvania	Tennessee	Ohio	Oklahoma
Rhode Island	Virginia*	South Dakota	Oregon
Vermont	West Virginia	Wisconsin	Texas
Virginia*			Utah
			Washington
			Wyoming

* The part of Virginia that is included in the Washington, DC, metropolitan area is included in the Northeast region; the remainder of the state is in the Southeast region.

[2] For the State Assessment Program, a PSU is most often a single school; for the national assessment, a PSU is a selected geographic region (a county, group of counties, or a metropolitan statistical area).

Gender

Results are reported separately for males and females.

Race/Ethnicity

The race/ethnicity variable is derived from two questions asked of students and schools' records, and it is used for race/ethnicity subgroup comparisons. Two questions from the set of general student background questions were used to determine race/ethnicity:

If you are Hispanic, what is your Hispanic background?

- I am not Hispanic.
- Mexican, Mexican American, or Chicano
- Puerto Rican
- Cuban
- Other Spanish or Hispanic background

Students who responded to this question by filling in the second, third, fourth, or fifth oval were considered Hispanic. For students who filled in the first oval, did not respond to the question, or provided information that was illegible or could not be classified, responses to the question below were examined in an effort to determine race/ethnicity.

Which best describes you?

- White (not Hispanic)
- Black (not Hispanic)
- Hispanic ("Hispanic" means someone who is from a Mexican, Mexican American, Chicano, Puerto Rican, Cuban, or other Spanish or Hispanic background.)
- Asian or Pacific Islander ("Asian or Pacific Islander" means someone who is from a Chinese, Japanese, Korean, Filipino, Vietnamese, or other Asian or Pacific Island background.)
- American Indian or Alaskan Native ("American Indian or Alaskan Native" means someone who is from one of the American Indian tribes, or one of the original people of Alaska.)
- Other (specify) _____

Students' race/ethnicity was then assigned on the basis of their response. For students who filled in the sixth oval ("Other") or provided illegible information or information that could not be classified, or did not respond at all, race/ethnicity was assigned as determined by school records.[1]

Race/ethnicity could not be determined for students who did not respond to either of the demographic questions and whose schools did not provide information about race/ethnicity.

The details of how race/ethnicity classifications were derived is presented so that readers can determine how useful the results are for their particular purposes. Also, some students indicated that they were from a Hispanic background (e.g., Puerto Rican or Cuban) and that a racial/ethnic category other than Hispanic best described them. These students were classified as Hispanic based on the rules described above. Furthermore, information from the schools did not always correspond to how students described themselves. Therefore, the racial/ethnic results presented in this report attempt to provide a clear picture based on several sources of information.

Parents' Highest Level of Education

The variable representing level of parental education is derived from responses to two questions from the set of general student background questions. Students were asked to indicate the extent of their mother's education:

How far in school did your mother go?

- She did not finish high school.
- She graduated from high school.
- She had some education after high school.
- She graduated from college.
- I don't know.

Students were asked a similar question about their father's education level:

How far in school did your father go?

- He did not finish high school.
- He graduated from high school.
- He had some education after high school.
- He graduated from college.
- I don't know.

[1] The procedure for assigning race/ethnicity was modified for Hawaii. See the *Technical Report for the NAEP 1996 State Assessment Program in Mathematics* for details.

The information was combined into one parental education reporting variable determined through the following process. If a student indicated the extent of education for only one parent, that level was included in the data. If a student indicated the extent of education for both parents, the higher of the two levels was included in the data. If a student did not know the level of education for both parents or did not know the level for one parent and did not respond for the other, the parental education level was classified as "I don't know." If the student did not respond for either parent, the student was recorded as having provided no response. (Nationally, 36 percent of fourth graders and 11 percent of eighth graders reported that they did not know the education level of either of their parents.)

Type of Location

Results are provided for students attending public schools in three mutually exclusive location types — central city, urban fringe/large town, and rural/small town — as defined below. The type of location variable is defined in such a way as to indicate the *geographical location* of a student's school. The intention is not to indicate, or imply, social or economic meanings for these location types. The type of location variable, on which the current NAEP sampling is based, does not support the reporting of regional results. Therefore, only state and national results will be presented.

Central City: The Central City category includes central cities of all Metropolitan Statistical Areas (MSAs).[4] Central City is a geographic term and is not synonymous with "inner city."

Urban Fringe/Large Town: An Urban Fringe includes all densely settled places and areas within MSAs that are classified as urban by the Bureau of the Census. A Large Town is defined as places outside MSAs with a population greater than or equal to 25,000.

Rural/Small Town: Rural includes all places and areas with a population of less than 2,500 that are classified as rural by the Bureau of the Census. A Small Town is defined as places outside MSAs with a population of less than 25,000 but greater than or equal to 2,500.

[4] Each Metropolitan Statistical Area (MSA) is defined by the Office of Management and Budget.

Type of School

Samples for the 1996 state assessment program were expanded to include students attending nonpublic schools (Catholic schools and other religious and private schools) in addition to students attending public schools. The expanded coverage was instituted for the first time in 1994. Samples for the 1990 and 1992 Trial State Assessment programs had been restricted to public school students only. For those jurisdictions meeting pre-established participation rate standards (see earlier section of this appendix), separate results are reported for public schools, for nonpublic schools, and for the combined public and nonpublic school samples. The combined sample for each jurisdiction also contains students attending Bureau of Indian Affairs (BIA) schools and Department of Defense Domestic Dependent Elementary and Secondary Schools (DDESS) in that jurisdiction. These two categories of schools are not included in either the public or nonpublic school samples.

Note that the DDESS and Department of Defense Dependents Schools (DoDDS)[5] were assessed in 1996 as separate jurisdictions, reported as jurisdictions with public school samples only.

Title I Participation

Based on available school records, students were classified as either currently participating in a Title I program or receiving Title I services, or as not receiving such services. The classification applies only to the school year when the assessment was administered (i.e., the 1995-96 school year) and is not based on participation in previous years. If the school did not offer any Title I programs or services, all students in that school were classified as not participating.

Eligibility for the Free/Reduced-Price School Lunch Program

Based on available school records, students were classified as either currently eligible for the free/reduced-price lunch component of the Department of Agriculture's National School Lunch Program or not eligible. The classification refers only to the school year when the assessment was administered (i.e., the 1995-96 school year) and is not based on eligibility in previous years. If school records were not available, the student was classified as "Information not available." If the school did not participate in the program, all students in that school were classified as "Information not available."

A.3 Guidelines for Analysis and Reporting

This report describes mathematics performance for eighth graders and compares the results for various groups of students within these populations — for example, those who have certain demographic characteristics or who responded to a specific background question in a particular way. The report examines the results for individual demographic groups and individual background questions. It does not include an analysis of the relationships among combinations of these subpopulations or background questions.

[5] The Department of Defense Dependents Schools (DoDDS) refers to overseas schools (i.e., schools outside the United States). Department of Defense Domestic Dependent Elementary and Secondary Schools (DDESS) refers to domestic schools (i.e., schools in the United States).

Drawing Inferences from the Results

Because the percentages of students in these subpopulations and their average scale scores are based on samples — rather than on the entire population of eighth graders in a jurisdiction — the numbers reported are necessarily *estimates*. As such, they are subject to a measure of uncertainty, reflected in the *standard error* of the estimate. When the percentages or average scale scores of certain groups are compared, it is essential to take the standard error into account, rather than to rely solely on observed similarities or differences. Therefore, the comparisons discussed in this report are based on *statistical tests* that consider both the magnitude of the difference between the averages or percentages and the standard errors of those statistics.

One of the goals of the state assessment program is to estimate scale score distributions and percentages of students in the categories described in A.2 for the overall populations of fourth- and eighth-grade students in each participating jurisdiction based on the particular samples of students assessed. The use of *confidence intervals*, based on the standard errors, provides a way to make inferences about the population average scale scores and percentages in a manner that reflects the uncertainty associated with the sample estimates. An estimated sample average scale score ± 2 standard errors approximates a *95 percent confidence interval* for the corresponding population average or percentage. This means that one can conclude with approximately 95 percent confidence that the average scale score of the entire population of interest (e.g., all fourth-grade students in public schools in a jurisdiction) is within ± 2 standard errors of the sample average.

As an example, suppose that the average mathematics scale score of the students in a particular jurisdiction's eighth-grade sample were 256 with a standard error of 1.2. A 95 percent confidence interval for the population average would be as follows:

$$\text{Mean} \pm 2 \text{ standard errors} = 256 \pm 2 \times (1.2) = 256 \pm 2.4 =$$

$$256 - 2.4 \text{ and } 256 + 2.4 = (253.6, 258.4)$$

Thus, one can conclude with 95 percent confidence that the average scale score for the entire population of eighth-grade students in public schools in that jurisdiction is between 253.6 and 258.4.

Similar confidence intervals can be constructed for percentages, *if the percentages are not extremely large or extremely small*. For extreme percentages, confidence intervals constructed in the above manner may not be appropriate, and accurate confidence intervals can be constructed only by using procedures that are quite complicated.

Extreme percentages, defined by both the magnitude of the percentage and the size of the sample from which it was derived, should be interpreted with caution. (The forthcoming *Technical Report of the NAEP 1996 State Assessment Program in Mathematics* contains a more complete discussion of extreme percentages.)

Analyzing Subgroup Differences in Averages and Percentages

The statistical tests determine whether the evidence — based on the data from the groups in the *sample* — is strong enough to conclude that the averages or percentages are really different for those groups in the *population*. If the evidence is strong (i.e., the difference is statistically significant), the report describes the group averages or percentages as being different (e.g., one group performed *higher than* or *lower than* another group) — regardless of whether the sample averages or sample percentages appear to be about the same or not. If the evidence is not sufficiently strong (i.e., the difference is not statistically significant), the averages or percentages are described as being *not significantly different* — again, regardless of whether the sample averages or sample percentages appear to be about the same or widely discrepant. The reader is cautioned to rely on the results of the statistical tests — rather than on the apparent magnitude of the difference between sample averages or percentages — to determine whether those sample differences are likely to represent actual differences between the groups in the population.

In addition to the overall results, this report presents outcomes separately for a variety of important subgroups. Many of these subgroups are defined by shared characteristics of students, such as their gender or race/ethnicity and the type of location in which their school is situated. Other subgroups are defined by the responses of the assessed students' mathematics teachers to questions in the mathematics teacher questionnaire.

In Chapter 1 of this report, differences between the jurisdiction and the nation were tested for overall mathematics scale score and for each of the mathematics content areas. In Chapter 2, significance tests were conducted for the overall scale score for each of the subpopulations. Chapter 3 reports differences between the jurisdiction and nation for the percentage of students at or above the *Proficient* level, and Chapter 4 contains significance tests for the percentage of students at or above the *Proficient* level for each of the subpopulations. In Chapters 5 through 7, comparisons were made across subgroups for responses to various background questions.

As an example of comparisons across subgroups, consider the question: *Do students who reported discussing studies at home almost every day exhibit higher average mathematics scale scores than students who report never or hardly ever doing so?*

To answer the question posed above, begin by comparing the average mathematics scale score for the two groups being analyzed. If the average for the group that reported discussing their studies at home almost every day is higher, it may be tempting to conclude that that group does have a higher mathematics scale score than the group that reported never or hardly ever discussing their studies at home. However, even though the averages differ, there may be no real difference in performance between the two groups in the population because of the uncertainty associated with the estimated average scale scores of the groups in the sample. Remember that the intent is to make a statement about the entire population, not about the particular sample that was assessed. The data from the sample are used to make inferences about the population as a whole.

As discussed in the previous section, each estimated sample average scale score (or percentage) has a degree of uncertainty associated with it. It is therefore possible that if all students in the population (rather than a sample of students) had been assessed or if the assessment had been repeated with a different sample of students or a different, but equivalent, set of questions, the performances of various groups would have been different. Thus, to determine whether there is a *real* difference between the average scale score (or percentage of a certain attribute) for two groups in the population, an estimate of the degree of uncertainty associated with the difference between the scale score averages or percentages of those groups must be obtained for the sample. This estimate of the degree of uncertainty — called *the standard error of the difference* between the groups — is obtained by taking the square of each group's standard error, summing these squared standard errors, and then taking the square root of this sum.

In a manner similar to that in which the standard error for an individual group average or percentage is used, the *standard error of the difference* can be used to help determine whether differences between groups in the population are real. The difference between the mean scale score or percentage of the two groups — 2 *standard errors of the difference* — represents an approximate 95 percent confidence interval. If the resulting interval includes zero, there is insufficient evidence to claim a real difference between groups in the population. If the interval does not contain zero, the difference between groups is *statistically significant* (different) at the .05 level.

As another example, to determine whether the average mathematics scale score of fourth-grade males is higher than that of fourth-grade females in a particular jurisdiction's public schools, suppose that the sample estimates of the average scale scores and standard errors for males and females were as follows:

Group	Average Scale Score	Standard Error
Males	218	0.9
Females	216	1.1

The difference between the estimates of the average sale scores of males and females is two points (218 - 216). The standard error of this difference is

$$\sqrt{0.9^2 + 1.1^2} = 1.4$$

Thus, an approximate 95 percent confidence interval for this difference is

Mean difference ± 2 standard errors of the difference =

$$2 \pm 2 \times (1.4) = 2 \pm 2.8 = 2 - 2.8 \text{ and } 2 + 2.8 = (-0.8, 4.8)$$

The value zero is within this confidence interval, which extends from -0.8 to 4.8 (i.e., zero is between -0.8 and 4.8). Thus, there is insufficient evidence to claim a difference in average mathematics scale score between the populations of fourth-grade males and females in public schools in the hypothetical jurisdiction.[6]

Throughout this report, when the average scale scores or percentages for two groups were compared, procedures like the one described above were used to draw the conclusions that are presented. If a statement appears in the report indicating that a particular group had a *higher* (or *lower*) average scale score than a second group, the 95 percent confidence interval for the difference between groups did not contain zero. An attempt was made to distinguish between group differences that were statistically significant but rather small in a practical sense and differences that were both statistically and practically significant. A procedure based on effect sizes was used. Statistically significant differences that are rather small are described in the text as *somewhat higher* or *somewhat lower*. When a statement indicates that the average scale score or percentage of some attribute was *not significantly different* for two groups, the confidence interval included zero, and thus no difference could be assumed between the groups. The information described in this section also pertains to comparisons across years. The reader is cautioned to avoid drawing conclusions solely on the basis of the magnitude of the difference. A difference between two groups in the sample that appears to be slight may represent a statistically significant difference in the population because of the magnitude of the standard errors. Conversely, a difference that appears to be large may not be statistically significant.

The procedures described in this section, and the certainty ascribed to intervals (e.g., a 95% confidence interval), are based on statistical theory that assumes that only one confidence interval or test of statistical significance is being performed. However, in each chapter of this report, many different groups are being compared (i.e., multiple sets of confidence intervals are being calculated). In sets of confidence intervals, statistical theory indicates that the certainty associated with the entire set of intervals is less than that attributable to each individual comparison from the set. To hold the certainty level for the set of comparisons at a particular level (e.g., 0.95), adjustments (called multiple comparison procedures) must be made to the methods described in the previous section. One such procedure — the *Bonferroni* method — was used in the analyses described in this report to form confidence intervals for the differences between groups whenever sets of comparisons were considered.[7] Thus, the confidence intervals in the text that are based on sets of comparisons are more conservative than those described on the previous pages.

---

[6] The procedure described above (especially the estimation of the standard error of the difference) is, in a strict sense, only appropriate when the statistics being compared come from independent samples. For certain comparisons in the report, the groups were not independent. In those cases, a different (and more appropriate) estimate of the standard error of the difference was used.

[7] Miller, R.G. *Simultaneous Statistical Inference*. (New York, NY: Wiley, 1966).

Most of the multiple comparisons in this report pertain to relatively small sets or "families" of comparisons. For example, when comparisons were discussed concerning students' reports of parental education, six comparisons were conducted — all pairs of the four parental education levels. In these situations, Bonferroni procedures were appropriate. However, the maps in Chapter 1 of this report display comparisons between New Hampshire and all other participating jurisdictions. The "family" of comparisons in this case was as many as 46. To control the certainty level for a large family of comparisons, the False Discovery rate (FDR) criterion[8] was used. Unlike the Bonferroni procedures which control the familywise error rate (i.e., the probability of making even one false rejection in the set of comparisons), the Benjamini and Hochberg (BH) approach using the FDR criterion controls the expected proportion of falsely rejected hypotheses as a proportion of all rejected hypotheses. Bonferroni procedures may be considered conservative for large families of comparisons.[9] In other words, using the Bonferroni method would produce more statistically nonsignificant comparisons than using the BH approach. Therefore, the BH approach is potentially more powerful for comparing New Hampshire to all other participating jurisdictions. A more detailed description of the Bonferroni and BH procedures appears in the *Technical Report of the NAEP 1996 State Assessment Program in Mathematics*.

Statistics with Poorly Estimated Standard Errors

Not only are the averages and percentages reported in NAEP subject to uncertainty, but their standard errors are as well. In certain cases, typically when the standard error is based on a small number of students or when the group of students is enrolled in a small number of schools, the amount of uncertainty associated with the standard errors may be quite large. Throughout this report, estimates of standard errors subject to a large degree of uncertainty are followed by the symbol "!". In such cases, the standard errors — and any confidence intervals or significance tests involving these standard errors — should be interpreted cautiously. Further details concerning procedures for identifying such standard errors are discussed in the *Technical Report of the NAEP 1996 State Assessment Program in Mathematics*.

Minimum Subgroup Sample Sizes

Results for mathematics performance and background variables were tabulated and reported for groups defined by gender, race/ethnicity, parental education, location of the school, type of school, participation in federally funded Title I programs, and eligibility for the free/reduced-price lunch component of the National School Lunch Program. NAEP collects data for five racial/ethnic subgroups (White, Black, Hispanic, Asian/Pacific Islander, and American Indian/Alaskan Native), three types of locations (Central City, Urban Fringe/Large Town, and Rural/Small Town),[10] and five levels of parents' education (Graduated From College, Some Education After High School, Graduated From High School, Did Not Finish High School, and I Don't Know).

[8] Benjamini, Y. and Y. Hochberg. "Controlling the false discovery rate: A practical and powerful approach to multiple testing," in *Journal of the Royal Statistical Society, Series B*, 57(1). (pp. 289—300), 1994).

[9] Williams, V.S.L., L.V. Jones, and J.W. Tukey. *Controlling Error in Multiple Comparisons, with Special Attention to the National Assessment of Educational Progress*. (Research Triangle Park, NC: National Institute of Statistical Sciences, December 1994).

[10] Previous NAEP reports reported data for four types of communities, rather than for the three types of location. These types of communities were Advantaged Urban, Disadvantaged Urban, Extreme Rural, and Other types of communities.

In many jurisdictions, and for some regions of the country, the number of students in some of these groups was not sufficiently high to permit accurate estimation of performance and/or background variable results. As a result, data are not provided for the subgroups with students from very few schools or for the subgroups with very small sample sizes. For results to be reported for any state assessment program subgroup, public school results must represent at least 5 primary sampling units (PSUs) and nonpublic school results must represent 6 schools. For results to be reported for any national assessment subgroup, at least 5 PSUs must be represented in the subgroup. In addition, a minimum sample of 62 students per subgroup is required. For statistical tests pertaining to subgroups, the sample size for both groups has to meet the minimum sample size requirements.

The minimum sample size of 62 was determined by computing the sample size required to detect an effect size of 0.5 total-group standard deviation units with a probability of 0.8 or greater. The effect size of 0.5 pertains to the *true* difference between the average scale score of the subgroup in question and the average scale score for the total fourth- or eighth-grade public school population in the jurisdiction, divided by the standard deviation of the scale score in the total population. If the *true* difference between subgroup and total group mean is 0.5 total-group standard deviation units, then a sample size of at least 62 is required to detect such a difference with a probability of 0.8. Further details about the procedure for determining minimum sample size appear in the *Technical Report of the NAEP 1996 State Assessment Program in Mathematics*.

Describing the Size of Percentages

Some of the percentages reported in the text of the report are given qualitative descriptions. For example, the number of students currently taking an algebra class might be described as "relatively few" or "almost all," depending on the size of the percentage in question. Any convention for choosing descriptive terms for the magnitude of percentages is to some degree arbitrary. The descriptive phrases used in the report and the rules used to select them are shown below.

Percentage	Descriptive Term Used in Report
$p = 0$	None
$0 < p \leq 8$	A small percentage
$8 < p \leq 13$	Relatively few
$13 < p \leq 18$	Less than one fifth
$18 < p \leq 22$	About one fifth
$22 < p \leq 27$	About one quarter
$27 < p \leq 30$	Less than one third
$30 < p \leq 36$	About one third
$36 < p \leq 47$	Less than half
$47 < p \leq 53$	About half
$53 < p \leq 64$	More than half
$64 < p \leq 71$	About two thirds
$71 < p \leq 79$	About three quarters
$79 < p \leq 89$	A large majority
$89 < p < 100$	Almost all
$p = 100$	All

APPENDIX B

The NAEP 1996 Mathematics Assessment

The 1996 assessment was the first update of the NAEP mathematics assessment framework[1] since the release of the National Council of Teachers of Mathematics (NCTM) *Curriculum and Evaluation Standards for School Mathematics*.[2] This update reflected refinements in the specifications governing the development of the 1996 assessment while assuring comparability of results across the 1990, 1992, and 1996 assessments. The refinements that distinguish the framework of the assessment conducted in 1996 from the framework of the assessments conducted in 1990 and 1992 include the following:

- moving away from the rigid content-strand-by-cognitive-process matrix that governed the development of earlier assessments. Classifying specific questions into cells of a matrix had required those questions to measure a unique content strand at a unique cognitive level. This stipulation often decontextualized the questions and limited the possibility of assessing students' abilities to reason in rich problem-solving situations and to make connections among content strands within mathematics.
- allowing individual questions on the assessment to be classified in one or more content strands when appropriate. Knowledge or skills from more than one content strand is often needed to answer a question. The option to classify questions in multiple ways provides a greater opportunity to measure student ability in content settings that closely approximate real-world reasoning and problem-solving situations. (However, to develop content strand scales, the primary content classification was used for questions with multiple classifications.)
- including the mathematics ability categories (conceptual understanding, procedural knowledge, and problem solving) as well as the process goals from the NCTM *Standards* (i.e., communication and connections) to achieve a balance of questions that measured a range of cognitive outcomes.

[1] National Assessment Governing Board. *Mathematics Framework for the 1996 National Assessment of Educational Progress*. (Washington, DC: National Assessment Governing Board, 1994).

[2] National Council of Teachers of Mathematics. *Curriculum and Evaluation Standards for School Mathematics*. (Reston, VA: NCTM, 1989).

- continuing the move towards including more constructed-response questions.

- creating "families" of questions that probe a student's understanding of mathematics vertically within a content strand or horizontally across content strands.

- revising the number sense, properties, and operations and geometry and spatial sense content strands to reflect the NCTM *Standards* emphasis on developing and assessing students' abilities to make sense of both number and operation and spatial settings.

These refinements to the NAEP mathematics framework were made so that the 1996 assessment would: (1) more adequately reflect recent curricular emphases and objectives and yet (2) maintain a connection with the 1990 and 1992 assessments to measure trends in student performance. Prior to the 1996 assessment, investigations were conducted to ensure that results from the assessment could be reported on the existing NAEP mathematics scale. The conclusion drawn from these investigations was that results from the 1990, 1992, and 1996 assessments could be reported on a common scale and trends in mathematics performance since 1990 examined.

The Assessment Design

Each student in the state assessment program in mathematics received a booklet containing a set of general background questions, a set of subject-specific background questions, and a combination of cognitive questions grouped in sets called blocks. At each grade level, the blocks of questions consisted of multiple-choice and constructed-response questions. Two types of constructed-response questions were included — short and extended constructed-response. Short constructed-response questions required students to provide answers to computation problems or to describe solutions in one or two sentences. Extended constructed-response questions required students to provide longer answers (e.g., a description of possibilities, a more involved computational analysis, or a description of a pattern and its implications). Students were expected to adequately answer the short constructed-response questions in about 2 to 3 minutes and the extended constructed-response questions in approximately 5 minutes. Short constructed-response questions which first appeared in the assessment in 1996 were graded to allow for partial credit (i.e., giving students credit for answers that are partially correct) according to a unique scoring rubric developed for each constructed-response question. Short constructed-response questions included in the 1990 and 1992 mathematics assessments were dichotomously scored (i.e., correct or incorrect). The extended constructed-response questions included in the 1992 and 1996 assessments were scored allowing for partial credit.

The blocks of questions contained several other features. Five to seven of the blocks at each grade level allowed calculator usage. At grade 4, students were provided four-function calculators, and at grade 8, students were provided scientific calculators. Prior to the assessment, all students were trained in the use of these calculators. For several blocks, students were given manipulatives (including geometric shapes, three-dimensional models, and spinners). For two of the blocks at each grade level, students were given rulers (at grade 4) or rulers and protractors (at grade 8) so the student could answer questions dealing with measurements and draw specified geometric shapes.

As part of the national assessment, other blocks of questions were developed for each of the grade levels. Each grade level had two estimation blocks that employed a paced-audiotape format to measure students' estimation skills. Each grade level also had two 30-minute theme blocks consisting of a mixture of multiple-choice and constructed-response questions. All of the questions in these blocks related to some aspect of a rich problem setting that served as a unifying theme for the entire block. Neither the estimation nor the theme block component were included in the state assessment program. Results for the estimation and theme blocks will be featured in future reports on the NAEP 1996 mathematics assessment.

Of the 17 blocks in the national sample at the fourth grade and the 19 blocks in the national sample at the eighth grade, 3 were carried forward from the 1990 assessment and 5 were carried forward from the 1992 assessment to allow for the measurement of trends across time. The remaining blocks of questions at each grade level contained new questions developed for the 1996 assessment as specified by the updated framework.

The data in Table B.1 reflect the number of questions by type by grade level for the 1990, 1992, and 1996 assessments. As mentioned earlier, the 1996 assessment continued NAEP's shift toward more constructed-response questions, including extended constructed-response questions that required students to provide an answer and a corresponding explanation.

THE NATION'S REPORT CARD 1996 State Assessment

TABLE B.1

Distribution of Questions by Question Type

	Grade 4			Grade 8			Grade 12		
	1990	1992	1996	1990	1992	1996	1990	1992	1996
Multiple-Choice	102	99	81	149	118	102	156	115	99
Short Constructed-Response*	41	59	64	42	65	69	47	64	74
Extended Constructed-Response**	---	5	13	---	6	12	---	6	11
Total	143	163	158	191	189	183	203	185	184

* Short constructed-response questions included in the 1990 and 1992 assessments were scored dichotomously. New short constructed-response questions included in the 1996 assessment were scored to allow for partial credit.
** No extended constructed-response questions were included in the 1990 assessment.

Each booklet in the state assessment program included three sets of student background questions. The first, consisting of general background questions, included questions about race or ethnicity, mother's and father's level of education, reading materials in the home, homework, attendance, and academic expectations. The second set, consisting of mathematics background questions, included questions about instructional activities, courses taken, use of specialized resources such as calculators in mathematics classes, and views on the utility and value of the subject. (Students were given 5 minutes to complete each set of questions, with the exception of the fourth graders, who were given more time because the general background questions were read aloud to them.) The third set of questions followed the cognitive question blocks and contained five questions about students' motivation to do well on the assessment, their perception of the difficulty of the assessment, and their familiarity with the types of cognitive questions included.

The blocks of cognitive and background questions were carefully balanced to ensure that the blocks could be completed within the time provided to the students, using information gathered from the field test. For more information on the design of the assessment, the reader is referred to Appendix C.

APPENDIX C

Technical Appendix: The Design, Implementation, and Analysis of the 1996 State Assessment Program in Mathematics

C.1 Overview

The purpose of this appendix is to provide technical information about the 1996 state assessment program in mathematics. It provides a description of the design for the assessment and gives an overview of the steps involved in the implementation of the program from the planning stages through to the analysis of the data.

This appendix is one of several documents that provide technical information about the 1996 state assessment program. Those interested in more details are referred to the forthcoming *Technical Report of the NAEP 1996 State Assessment Program in Mathematics*. Theoretical information about the models and procedures used in NAEP can be found in the special NAEP-related issue of the *Journal of Educational Statistics* (Summer 1992/Volume 17, Number 2) as well as previous national technical reports.

Educational Testing Service (ETS) was awarded the cooperative agreement for the 1996 NAEP programs, including the state assessment program. ETS was responsible for overall management of the programs as well as for development of the overall design, the cognitive questions and questionnaires, data analysis, and reporting. National Computer Systems (NCS) was a subcontractor to ETS on both the national and state NAEP programs. NCS was responsible for printing, distribution, and receipt of all assessment materials, and for scanning and professional scoring. All aspects of sampling and field operations for both the national and state assessment programs were the responsibility of Westat, Inc. NCES awarded a separate cooperative agreement to Westat for these services for the national and state assessments.

Organization of the Technical Appendix

This appendix provides a brief description of the design for the state assessment program in mathematics and gives an overview of the steps involved in implementing the program from the planning stages to the analysis of the data. (A more detailed discussion of the technical aspects of the NAEP state assessment program can be found in the forthcoming *Technical Report of the NAEP 1996 State Assessment Program in Mathematics*.) The organization of this appendix is as follows:

- Section C.2 provides an overview of the design of the 1996 state assessment program in mathematics.
- Section C.3 discusses the balanced incomplete block (BIB) spiral design that was used to assign cognitive questions to assessment booklets and assessment booklets to students.
- Section C.4 outlines the sampling design used for the 1996 state assessment program.
- Section C.5 summarizes Westat's field administration procedures.
- Section C.6 describes the flow of the data from their receipt at NCS through data entry and professional scoring.
- Section C.7 summarizes the procedures used to weight the assessment data and to obtain estimates of the sampling variability of subpopulation estimates.
- Section C.8 describes the initial analyses performed to verify the quality of the data.
- Section C.9 describes the item response theory scales and the overall mathematics composite scale that were created for the final analyses of the state assessment program data.
- Section C.10 provides an overview of the linking of the scaled results from the state assessment program in mathematics to those from the national assessment.

C.2 Design of the NAEP 1996 State Assessment Program in Mathematics

The major aspects of the design for the state assessment program in mathematics included the following:

- Participation at the jurisdiction level was voluntary.
- Fourth- and eighth-grade students from public and nonpublic schools were assessed. Nonpublic schools included Catholic schools, other religious schools, private schools, Department of Defense Domestic Elementary and Secondary Schools (DDESS), and Bureau of Indian Affairs schools. Separate representative samples of public and nonpublic schools were selected in each participating jurisdiction and students were randomly sampled within schools. The size of a jurisdiction's nonpublic school samples was proportional to the percentage of students in that jurisdiction attending such schools.

- The fourth- and eighth-grade mathematics assessment instruments used for the state assessment program and the national assessment consisted of 13 blocks of questions. Eight of these blocks were previously administered as part of the 1990 and 1992 national and Trial State Assessments. The type of questions — constructed-response or multiple-choice — was determined by the nature of the task. In addition, the constructed-response questions were of two types: *short constructed-response* questions required students to provide answers to computation problems or to describe solutions in one or two sentences, while *extended constructed-response* questions required students to provide longer responses when answering the question. Each student was given 3 of the 13 blocks of questions.

- A complex form of matrix sampling called a balanced incomplete block (BIB) spiraling design was used. With BIB spiraling, students in an assessment session received different booklets, which provided for greater mathematics content coverage than would have been possible had every student been administered the identical set of questions, without imposing an undue testing burden on the student.

- Background questionnaires given to the students, the students' mathematics teachers, and the principals or other administrators provided a variety of contextual information. The background questionnaires for the state assessment program were identical to those used in the national fourth- and eighth-grade assessments.

- The total assessment time for each student was approximately one hour and 40 minutes. Each assessed student was assigned a mathematics booklet that contained two 5-minute background questionnaires, followed by 3 of the 13 blocks of mathematics questions requiring 15 minutes each, and a 3-minute motivation questionnaire. Twenty-six different booklets were assembled.

- The assessments were scheduled to take place in the five-week period between January 29 and March 4, 1996. One-fourth of the schools in each jurisdiction were to be assessed each week throughout the first four weeks; however, due to the severe weather throughout much of the country, the fifth week was used for regular testing as well as for makeup sessions.

- Data collection was, by law, the responsibility of each participating jurisdiction. Security and uniform assessment administration were high priorities. Extensive training of state assessment personnel was conducted to assure that the assessment would be administered under standard, uniform procedures. For jurisdictions that had participated in previous NAEP state assessments, 25 percent of both public and nonpublic school assessment sessions were monitored by the Westat staff. For the jurisdictions new to NAEP, 50 percent of both public and nonpublic school sessions were monitored.

C.3 Assessment Instruments

The assembly of cognitive questions into booklets and their subsequent assignment to assessed students was determined by a BIB design with spiraled administration. This design is a variant of a matrix sampling design. The full set of mathematics questions was divided into 13 unique blocks, each requiring 15 minutes for completion. Each assessed student received a booklet containing 3 of the 13 blocks according to a design that ensured that each block was administered to a representative sample of students within each jurisdiction.

In addition to the student assessment booklets, three other instruments provided data relating to the assessment — a mathematics teacher questionnaire, a school characteristics and policies questionnaire, and an SD/LEP student questionnaire.

The *student assessment booklets* contained five sections and included both cognitive and noncognitive questions. In addition to three 15-minute sections of cognitive questions, each booklet included two 5-minute sets of general and mathematics background questions designed to gather contextual information about students, their experiences in mathematics, and their attitudes toward the subject, and one 3-minute section of motivation questions designed to gather information about the student's level of motivation while taking the assessment.

The *teacher questionnaire* was administered to the mathematics teachers of the fourth- and eighth-grade students participating in the assessment. The questionnaire consisted of three sections and took approximately 20 minutes to complete. The first section focused on the teacher's general background and experience; the second, on the teacher's background related to mathematics; and the third, on classroom information about mathematics instruction.

The *school characteristics and policies questionnaire* was given to the principal or other administrator in each participating school and took about 20 minutes to complete. The questions asked about the principal's background and experience, school policies, programs, and facilities, and the demographic composition and background of the students and teachers.

The *SD/LEP student questionnaire* was completed by the staff member most familiar with any student selected for the assessment who was classified in either of two ways: students with disabilities (SD) had an Individualized Education Plan (IEP) of equivalent special education plan (for reasons other than being gifted and talented); students with limited English proficiency were classified as LEP students. The questionnaire took approximately three minutes to complete and asked about the student and the special programs in which the student participated. It was completed for all selected SD or LEP students regardless of whether or not they participated in the assessment. Selected SD or LEP students participated in the assessment if they were determined by the school to be able to participate, considering the terms of their IEP and accommodations provided by the school or by NAEP.

C.4 The Sampling Design

The sampling design for NAEP is complex, in order to minimize burden on schools and students while maximizing the utility of the data; for further details see the forthcoming *Technical Report for the NAEP 1996 State Assessment Program in Mathematics*. The target populations for the state assessment program in mathematics consisted of fourth- and eighth-grade students enrolled in either public or nonpublic schools. The representative samples of public school fourth and eighth graders assessed in the state assessment program came from about 100 schools (per grade) in most jurisdictions. However, if a jurisdiction had fewer than 100 public schools with a particular grade, all or almost all schools were asked to participate. If a jurisdiction had smaller numbers of students in each school than expected, more than 100 schools were selected for participation. The nonpublic school samples differed in size across the jurisdictions, with the number of schools selected proportional to the nonpublic school enrollment within each jurisdiction. Typically, about 20 to 25 nonpublic schools (per grade) were included for each jurisdiction. The school sample in each jurisdiction was designed to produce aggregate estimates for the jurisdiction and for selected subpopulations (depending upon the size and distribution of the various subpopulations within the jurisdiction) and also to enable comparisons to be made, at the jurisdiction level, between administration of assessment tasks with monitoring and without monitoring. The public schools were stratified by urbanization, percentage of Black and Hispanic students enrolled, and median household income within the ZIP code area of the school. The nonpublic schools were stratified by type of control (Catholic, private/other religious, other nonpublic), metropolitan status, and enrollment size per grade.

The national and regional results presented in this report are based on nationally representative samples of fourth- and eighth-grade students. The samples were selected using a complex multistage sampling design involving the sampling of students from selected schools within selected geographic areas across the country. The sample design had the following stages:

(1) selection of geographic areas (a county, group of counties, or metropolitan statistical area)

(2) selection of schools (public and nonpublic) within the selected areas

(3) selection of students within selected schools

Each selected school that participated in the assessment, and each student assessed, represent a portion of the population of interest. To make valid inferences from student samples to the respective populations from which they were drawn, sampling weights are needed. Discussions of sampling weights and how they are used in analyses are presented in sections C.7 and C.8.

The state results provided in this report are based on state-level samples of fourth- and eighth-grade students. The samples of both public and nonpublic school students were selected based on a two-stage sample design that entailed selecting students within schools. The first-stage samples of schools were selected with a probability proportional to the fourth- or eighth-grade enrollment in the schools. Special procedures were used for jurisdictions with many small schools and for jurisdictions with a small number of schools. As with the national samples, the state samples were weighted to allow for valid inferences about the populations of interest.

The results presented for a particular jurisdiction are based on the representative sample of students who participated in the 1996 state assessment program. The results for the nation and regions of the country are based on the nationally and regionally representative samples of students who were assessed as part of the national NAEP program. Using the national and regional results from the 1996 national assessment was necessary because of the voluntary nature of the state assessment program. Because not every state participated in the program, the aggregated data across states did not necessarily provide representative national or regional results.

In most jurisdictions, up to 30 students were selected from each school, with the aim of providing an initial sample size of approximately 3,000 public school students per jurisdiction per grade. The student sample size of 30 for each school was chosen to ensure that at least 2,000 public school students (per grade) participated from each jurisdiction, allowing for school nonresponse, exclusion of students, inaccuracies in the measures of enrollment, and student absenteeism from the assessment. In jurisdictions with fewer schools, larger numbers of students per school were often required to ensure initial samples of roughly 3,000 students. In certain jurisdictions, all eligible fourth or eighth graders were targeted for assessment. Jurisdictions were given the option to reduce the expected student sample size in order to reduce testing burden and the number of multiple-testing sessions for participating schools. At grade 4, two jurisdictions (Delaware and Guam) and at grade 8, four jurisdictions (Alaska, Delaware, Hawaii, and Rhode Island) elected to exercise this option. Using this option can involve compromises such as higher standard errors and accompanying loss of precision.

In order to provide for wider inclusion of students with disabilities and limited English proficiency, the 1996 state assessments in mathematics involved dividing the sample of students at each grade level into two subsamples, referred to as S1 and S2. S1 provided continuity with the 1992 mathematics assessment and thus allowed for the reporting of performance over time by using the same exclusion criteria for students with disabilities and limited English proficiency as was used in that assessment. S2 provided for wider inclusion of students with disabilities and limited English proficiency by incorporating new exclusion rules. For further discussion, see the *NAEP 1996 Mathematics Report Card*. The 1996 national assessment in mathematics involved an additional subsample, S3, in which accommodations were provided for certain students with disabilities or limited English proficiency, again in order to make NAEP more inclusive.

For both the national and state mathematics assessments, scaling and analysis procedures (discussed in sections C.8 to C.10) were applied to a combination of students from S1 and S2. Specifically, all assessed students from S1 were combined with those students from S2 who were **not** identified as SD or LEP. This combination of segments of the S1 and S2 subsamples provided for maximizing the use of available data while allowing for comparisons to the student population in the national sample. This combination, referred to as the "reporting sample," was the sample used in linking the state assessment to the national assessment (see Section C.10).

Additional analyses will be conducted on the national samples in order to study the effects of changing the exclusion rules and the presence of accommodations. Preliminary discussion can be found in the *NAEP 1996 Mathematics Report Card* and more detailed discussion will follow in future NAEP publications.

C.5 Field Administration

The administration of the 1996 program required collaboration between staff in the participating jurisdictions and schools and the NAEP contractors, especially Westat, the field administration contractor.

Each jurisdiction volunteering to participate in the 1996 state assessment program was asked to appoint a state coordinator as liaison between NAEP staff and the participating schools. In addition, Westat hired and trained a supervisor for each jurisdiction and six field managers, each of whom was assigned to work with groups of jurisdictions. The state supervisors were responsible for working with the state coordinators, overseeing assessment activities, training school district personnel to administer the assessment, and coordinating the quality-control monitoring efforts. Each field manager was responsible for working with the state coordinators of seven to eight jurisdictions and for the supervision of the state supervisors assigned to those jurisdictions. An assessment administrator was responsible for preparing for and conducting the assessment session in one or more schools. These individuals were usually school or district staff and were trained by Westat. Westat also hired and trained three to five quality control monitors in each jurisdiction. For jurisdictions that had previously participated in the state assessment program, 25 percent of the public and nonpublic school sessions were monitored. For jurisdictions new to the program, 50 percent of all sessions were monitored. The assessment sessions were conducted during a five-week period beginning in late January 1996.

C.6 Materials Processing, Professional Scoring, and Database Creation

Upon completion of each assessment session, school personnel shipped the assessment booklets and forms to NCS for professional scoring, entry into computer files, and checking. The files were then sent to ETS for creation of the database.

After NCS received all appropriate materials from a school, they were forwarded to the professional scoring area where the responses to the constructed-response question were evaluated by trained staff using guidelines prepared by ETS. Each constructed-response question had a unique scoring guide that defined the criteria to be used in evaluating students' responses. The extended constructed-response questions were evaluated with four- or five-level rubrics, and the short constructed-response questions first used in 1996 were rated according to three-level rubrics that permit partial credit to be given. Short constructed-response questions used previously were scored dichotomously (i.e., correct or incorrect).

For the national mathematics assessment and the state assessment program in mathematics, over 4.8 million constructed responses were scored. This figure includes rescoring to monitor inter-rater reliability and trend reliability. In other words, scoring reliability was calculated both within year (1996) and across years (1990, 1992, and 1996). The overall within-year percentages of agreement for the 1996 national within-year reliability samples were 96 percent at grade 4 and 96 percent at grade 8. The percentages of agreement across the assessment years for the national inter-year reliability samples were 96 percent (1990 to 1996) and 94 percent (1992 to 1996) at grade 4 and 95 percent (1990 to 1996) and 94 percent (1992 to 1996) at grade 8.

Data transcription and editing procedures were used to generate the disk and tape files containing various assessment information, including the sampling weights required to make valid statistical inferences about the population from which the state assessment program sample was drawn. Prior to analysis, the data from these files underwent a quality control check at ETS. The files were then merged into a comprehensive, integrated database.

C.7 Weighting and Variance Estimation

A complex sample design was used to select the students to be assessed in each of the participating jurisdictions. The properties of a sample from a complex design are very different from those of a simple random sample in which every student in the target population has an equal chance of selection and in which the observations from different sampled students can be considered to be statistically independent of one another. The properties of the sample from the complex state assessment program design were taken into account in the analysis of the assessment data.

One way that the properties of the sample design were addressed was by using sampling weights to account for the fact that the probabilities of selection were not identical for all students. These weights also included adjustments for school and student nonresponse. All population and subpopulation characteristics based on the state assessment program data used sampling weights in their estimation.

In addition to deriving appropriate estimates of population characteristics, it is essential to obtain appropriate measures of the degree of uncertainty of those statistics. One component of uncertainty results from sampling variability, which is a measure of the dependence of the results on the particular sample of students actually assessed. Because of the effects of cluster selection (schools are selected first, then students are selected within those schools), observations made on different students cannot be assumed to be independent of each other (and, in fact, are generally positively correlated). As a result, classical variance estimation formulas will produce incorrect results. Instead, a jackknife variance estimation procedure that takes the characteristics of the sample into account was used for all analyses.

Jackknife variance estimation provides a reasonable measure of uncertainty for any statistic based on values observed without error. Statistics such as the percentage of students correctly answering a given question meet this requirement, but other statistics based on estimates of student mathematics performance, such as the average mathematics scale score of a subpopulation, do not. Because each student typically responds to relatively few questions from a particular content strand (e.g., Algebra and Functions or Geometry and Spatial Sense) there exists a nontrivial amount of imprecision in the measurement of the scale score of a given student. This imprecision adds an additional component of variability to statistics based on estimates of individual scale scores.

C.8 Preliminary Data Analysis

After the computer files of student responses were received from NCS and merged into an integrated database, all cognitive and noncognitive questions were subjected to an extensive item analysis. For each question, this analysis yielded the number of respondents, the percentage of responses in each category, the percentage who omitted the question, the percentage who did not reach the question, and the correlation between the question score and the block score. In addition, the item analysis program provided summary statistics for each block, including a reliability (internal consistency) coefficient. These analyses were used to check the scoring of the questions, to verify the appropriateness of the difficulty level of the questions, and to check for speededness. The results were reviewed by knowledgeable project staff in search of aberrations that might signal unusual results or errors in the database.

The question and block-level analyses were done using rescaled versions of the final sampling weights provided by Westat (see Section C.7). The rescaling was carried out within each jurisdiction. The sum of the sampling weights for the public school students within each jurisdiction was constrained to be equal. The same transformation was then applied to the weights of the nonpublic school students in that jurisdiction. The sum of the weights for each of the DoDEA samples (i.e., DDESS and DoDDS) was constrained to be equal to the same value as the public school students in other jurisdictions. Use of rescaled weights does nothing to alter the value of statistics calculated separately within each jurisdiction. However, for statistics obtained from samples that combine students from different jurisdictions, use of the rescaled weights results in a roughly equal contribution of each jurisdiction's data to the final value of the estimate. Equal contribution of each jurisdiction's data to the results of the item response theory (IRT) scaling was viewed as a desirable outcome. The original final sampling weights provided by Westat were used in reporting.

Additional analyses comparing the data from the monitored sessions with those from the unmonitored sessions were conducted to determine the comparability of the assessment data from the two types of administrations. Differential item functioning (DIF) analyses were carried out using the national assessment data. DIF analyses identify questions that were differentially difficult for various subgroups, affording the opportunity to reexamine such questions with respect to their fairness and their appropriateness for inclusion in the scaling process.

C.9 Scaling the Assessment Questions

The primary analysis and reporting of the results from the state assessment program used item response theory (IRT) scale-score models. Scaling models quantify a respondent's tendency to provide correct answers to the domain of questions contributing to a scale as a function of a parameter called performance, estimated by a scale score. The scale scores can be viewed as a summary measure of performance across the domain of questions that make up the scale. Three distinct IRT models were used for scaling: 1) 3-parameter logistic models for multiple-choice questions; 2) 2-parameter logistic models for short constructed-response questions that were scored correct or incorrect; and 3) generalized partial credit models for short and extended constructed-response questions that were scored on a multipoint (i.e., greater than two levels) scale.

Five distinct scales were created for the state assessment program in mathematics to summarize fourth- and eighth-grade students' abilities according to the five defined content strands (Number Sense, Properties, and Operations; Measurement; Geometry and Spatial Sense; Data Analysis, Statistics, and Probability; and Algebra and Functions). These scales were defined identically to, but separately from, those used for the scaling of the national NAEP fourth- and eighth-grade mathematics data. Although the questions comprising each scale were identical to those used in the national assessment program, the item parameters for the state assessment program scales were estimated from combined public school data from the jurisdictions participating in the state assessment program.[1] Item parameter estimation was carried out on an item calibration subsample. The calibration subsample consisted of an approximately 25 percent sample of all available public school data. To ensure equal representation in the scaling process, each jurisdiction contributed the same number of students to the item calibration sample. Within each jurisdiction, 50 percent of the calibration sample was taken from monitored administrations and the other 50 percent came from unmonitored administrations.

The fit of the IRT model to the observed data was examined within each scale by comparing the estimates of the empirical item characteristic functions with the theoretic curves. For correct-incorrect questions, nonmodel-based estimates of the expected proportions of correct responses to each question for students with various levels of scale proficiency were compared with the fitted item response curve; for the short and extended partial-credit constructed-response questions, the comparisons were based on the expected proportions of students with various levels of scale proficiency who achieved each score level. In general, the question-level results were well fit by the scaling models.

[1] Schools from the DoDEA jurisdictions were not included in the item calibration sample.

Using the item parameter estimates, estimates of various population statistics were obtained for each jurisdiction. The NAEP methods use random draws ("plausible values") from estimated proficiency distributions for each student to compute population statistics. Plausible values are not optimal estimates of individual student proficiencies; instead, they serve as intermediate values to be used in estimating population characteristics. Under the assumptions of the scaling models, these population estimates will be consistent, in the sense that the estimates approach the model-based population values as the sample size increases, which would not be the case for population estimates obtained by aggregating optimal estimates of individual performance.

In addition to the plausible values for each scale, a composite of the five content strand scales was created as a measure of overall mathematics proficiency. This composite was a weighted average of the five mathematics scales in which the weights were proportional to the relative importance assigned to each content strand in the mathematics framework. The definition of the composite for the state assessment program was identical to that used for the national fourth- and eighth-grade mathematics assessments.

C.10 Linking the State Results to the National Results

A major purpose of the state assessment program was to allow each participating jurisdiction to compare its 1996 results with those for the nation as a whole and with those for the region of the country in which that jurisdiction is located. For meaningful comparisons to be made between each jurisdiction and the relevant national sample, results from these two assessments had to be expressed in terms of a similar system of scale units.

The results from the state assessment program were linked to those from the national assessment through linking functions determined by comparing the results for the aggregate of all students assessed in the state assessment program with the results for students of the matching grade within the National Linking Sample of the national NAEP. The National Linking Sample of the national NAEP for a given grade is a representative sample of the population of all grade-eligible public school students within the aggregate of 45 participating states and the District of Columbia. Guam and the two Department of Defense Education Activity (DoDEA) jurisdictions were not included in the aggregate. Specifically, the fourth- and eighth-grade National Linking Samples consist of all fourth- and eighth-grade students in public schools in the states and the District of Columbia who were assessed in the national cross-sectional mathematics assessment.

For each grade, a linear equating within each scale was used to link the results of the state assessment program to the national assessment. For each scale, the adequacy of the linear equating was evaluated by comparing the distribution of mathematics scale scores based on the aggregation of all assessed students at each grade from the participating states and the District of Columbia with the equivalent distribution based on the students in the National Linking Sample. In the estimation of these distributions, the students were weighted to represent the target population of public school students in the specified grade in the aggregation of the states and the District of Columbia. If a linear equating were adequate, the distribution for the aggregate of states and the District of Columbia and that for the National Linking Sample will have, to a close approximation, the same shape in terms of the skewness, kurtosis, and higher moments of the distributions. The only differences in the distributions allowed by linear equating are in the means and variances. Generally, this has been found to be the case.

Each mathematics content-strand scale was linked by matching the mean and standard deviation of the scale scores across all students in the state assessment (excluding Guam and the two DoDEA jurisdictions) to the corresponding scale mean and standard deviation across all students in the National Linking Sample.

APPENDIX D

Setting the Achievement Levels

Setting achievement levels is a test-centered method for setting standards on the NAEP assessment that identifies what students should know and should be able to do. The method depends on securing and summarizing a set of judgmental ratings of expectations for student educational performance on specific questions comprising the NAEP mathematics assessment. The NAEP mathematics scale is a numerical index of students' performance in mathematics ranging from 0 to 500. The three achievement levels — *Basic, Proficient,* and *Advanced* — are mapped onto the scale for each grade level assessed.

The NAEP mathematics achievement levels were set following the 1990 assessment and further refined following the 1992 assessment. In developing the threshold values for the levels, a broadly constituted panel of judges — including teachers (50%), non-teacher educators (20%), and the general public (noneducators)[1] (30%) — rated a grade-specific item pool using the policy definitions of the National Assessment Governing Board (NAGB) for *Basic, Proficient,* and *Advanced.* The policy definitions were operationalized by the judges in terms of specific mathematical skills, knowledge, and behaviors that were judged to be appropriate expectations for students in each grade and were in accordance with the current mathematics assessment framework. The policy definitions are as follows:

Basic

This level denotes partial mastery of the prerequisite knowledge and skills that are fundamental for proficient work at each grade.

Proficient

This level represents solid academic performance for each grade assessed. Students reaching this level have demonstrated competency over challenging subject matter and are well prepared for the next level of schooling.

Advanced

This higher level signifies superior performance beyond proficient grade-level mastery at each grade.

[1] Noneducators represented business, labor, government service, parents, and the general public.

The judges' operationalized definitions were incorporated into lists of descriptors that represent what borderline students should be able to do at each of the levels defined by policy. The purpose of having panelists develop their own operational definitions of the achievement levels was to ensure that all panelists would have a common understanding of borderline performances and a common set of content-based referents to use during the item-rating process.

The judges (24 at grade 4 and 22 at grade 8) each rated half of the questions in the NAEP pool in terms of the expected probability that a student at a borderline achievement level would answer the question correctly, based on the judges' operationalization of the policy definitions and the factors that influence question difficulty. To assist the judges in generating consistently scaled ratings, the rating process was repeated twice, with feedback. Information on consistency among different judges and on the difficulty of each question[2] was fed back into the first repetition (round 2), while information on consistency within each judge's set of ratings was fed back into the second repetition (round 3). The third round of ratings permitted the judges to discuss their ratings among themselves to resolve problematic ratings. The mean final rating of the judges aggregated across questions yielded the threshold values in the percent correct metric. These cut scores were then mapped onto the NAEP scale (which is defined and scored using item response theory, rather than percent correct) to obtain the scale scores for the achievement levels.[3] The judges' ratings, in both metrics, and their associated errors of measurement are shown below. NAGB accepted the panel's achievement levels and, for reporting purposes, set final cutpoints one standard error (a measure of consistency among the judges' ratings) below the mean levels.

THE NATION'S REPORT CARD NAEP 1996 State Assessment

FIGURE D.1

Cutpoints for Achievement Levels at Grades 4 and 8

Grade	Level	Mean Percent Correct (Round 3)	Scale Score*	Standard Error of Scale Score**
4	Basic	39	214	1.9
4	Proficient	65	249	4.1
4	Advanced	84	282	4.0
8	Basic	48	262	2.4
8	Proficient	71	299	5.7
8	Advanced	87	333	4.8

* Scale score is derived from a weighted average of the mean percent correct for multiple-choice and short constructed-response questions after both were mapped onto the NAEP scale.

** The standard error of the scale score is estimated from the difference in mean scale scores for the two equivalent subgroups of judges.

[2] Item difficulty estimates were based on a preliminary, partial set of responses to the national assessment.

[3] See Appendix A for a discussion of the technical errors that resulted in the reanalysis and rereporting of 1990 and 1992 mathematics achievement level results.

After the ratings were completed, the judges for each grade level reviewed the operationalized descriptions developed by the judges of the other grade levels as well as their own descriptions and defined achievement level descriptions that were generally acceptable to all three grade-group judges. However, the descriptions varied in format, sharpness of language, and degree of specificity of the statements. Therefore, another panel at a subsequent validation meeting improved the wording and modified the language of the achievement level descriptions to reflect more closely the terminology of the National Council of Teachers of Mathematics *Curriculum and Evaluation Standards for School Mathematics*.[4] The achievement level descriptions, though based on the 1992 NAEP pool, apply to the current assessment and will not change from assessment to assessment (that is, until the framework changes).

[4] National Council of Teachers of Mathematics. *Curriculum and Evaluation Standards for School Mathematics*. (Reston, VA: NCTM, 1989).

ACKNOWLEDGMENTS

This report is the culmination of the efforts of many individuals who contributed their considerable knowledge, experience, and creativity to the NAEP 1996 mathematics assessment. The NAEP 1996 mathematics state assessment was a collaborative effort among staff from the National Center for Education Statistics (NCES), the National Assessment Governing Board (NAGB), Educational Testing Service (ETS), Westat, Inc., and National Computer Systems (NCS). In addition, the program benefited from the contributions of hundreds of individuals at the state and local levels — governors, chief state school officers, state and district test directors, state coordinators, and district administrators — who provided their wisdom, experience, and hard work. Most importantly, NAEP is grateful to the over 239,000 students and the teachers and administrators in over 9,700 schools in 48 jurisdictions who made the assessment possible by contributing considerable amounts of time and effort.

The NAEP 1996 mathematics state assessment was funded through NCES, in the Office of Educational Research and Improvement of the U.S. Department of Education. The Commissioner of Education Statistics, Pascal D. Forgione, Jr., and the NCES staff — Sue Ahmed, Peggy Carr, Arnold Goldstein, Steven Gorman, Larry Ogle, Gary W. Phillips, Sharif Shakrani, Maureen Treacy — and Alan Vanneman of the Education Statistics Services Institute, worked closely and collegially with the authors to produce this report. The authors were also provided invaluable advice and guidance by the members of the National Assessment Governing Board and NAGB staff. In particular, the authors are indebted to Arnold Goldstein of NCES for his daily efforts to coordinate the activities of the many people who contributed to this report.

The NAEP project at ETS is housed in the Center for the Assessment of Educational Progress under the direction of Paul Williams. The NAEP 1996 assessments were directed by Stephen Lazer and John Mazzeo. Jeff Haberstroh directed the scoring operations for the 1996 mathematics assessment. Sampling and data collection activities were conducted by Westat under the direction of Rene Slobasky, Nancy Caldwell, Keith Rust, Debby Vivari, and Dianne Walsh. Printing, distribution, scoring, and processing activities were conducted by NCS under the direction of Brad Thayer, Patrick Bourgeacq, Charles Brungardt, Mathilde Kennel, Linda Reynolds, and Connie Smith.

The complex statistical and psychometric activities necessary to report results for the NAEP 1996 mathematics assessment were directed by Nancy Allen, John Barone, James Carlson, and Juliet Shaffer. John Mazzeo and Gene Johnson provided direction on several

critical psychometric issues. The analyses presented in this report were led by Frank Jenkins and Edward Kulick, with assistance from Hua Chang, Steve Wang, Xiaohui Wang, Hong Zhou, Jiahe Qian, Kate Pashley, David Freund, and Norma Norris.

Laura Jerry was responsible for the development and creation of the computer-generated reports, with assistance from Xiaohui Wang, Laura Jenkins, Phillip Leung, Inge Novatkoski, Bruce Kaplan, and Alfred Rogers. A large group of NAEP staff at ETS checked the data, text, and tables. Debbie Kline coordinated the technical appendices.

Many thanks are due to the comments and critical feedback of numerous reviewers, both internal and external to NCES and ETS. Important contributions were made by reviewers from academic institutions and education agencies: Bruce Brombacher of Upper Arlington (Ohio) Schools; Pasquale DeVito of the Rhode Island Department of Education, John Dossey of Illinois State University, Thomas Fisher of the Florida Department of Education, Douglas Rindone of the Connecticut Department of Education, and Irvin Vance of Michigan State University. Valuable input was given by NAGB staff Mary Lynn Bourque and Lawrence Feinberg, and NCES staff Susan Ahmed, Peggy Carr (who helped guide the report through several versions), Steven Gorman, Andrew Kolstad, Mary Frase, Mary Rollefson, Sharif Shakrani, and Shi-Chang Wu.

Cover design and production of the print version was directed by Carol Errickson, with the assistance of Sharon Davis-Johnson, Alice Kass, and Barbette Tardugno. Karen Damiano produced tables and text for one state for which the computerized report generating system was not appropriate. The World Wide Web version of the state reports was produced by Phillip Leung and Pat O'Reilly with assistance from Debbie Kline, Karen Damiano, Sharon M. Davis-Johnson, Craig Pizzuti, Barbette Tardugno, and Christine Zelenak. The *NAEP 1996 Mathematics State Report* for all participating jurisdictions is available via **http://www.ed.gov/NCES/naep**.

★ U.S. GOVERNMENT PRINTING OFFICE: 1997 — 427 - 334 / 70390

NCES 97-974 NH